

Dipartimento di / Department of

Department of Informatics, Systems and Communication

Dottorato di Ricerca in / PhD program Computer Science Ciclo / Cycle XXXV

Curriculum in (se presente / if it is) _____

Raising Teenagers' Awareness of Social Media Threats: A Theoretical and Empirical Study

Cognome / Surname Lomonaco Nome / Name Francesco

Matricola / Registration number 876570

Tutore / Tutor: Fabio Stella

Cotutore / Co-tutor: _____
(se presente / if there is one)

Supervisor: Dimitri Ognibene
(se presente / if there is one)

Coordinatore / Coordinator: Leonardo Mariani

ANNO ACCADEMICO / ACADEMIC YEAR 2022/2023

Great disorder under the Heavens and the situation is excellent.

Mao Zedong

Se qualcuno ancora si commuove per questa generazione,
non c'ha chiaro il punto della situazione

Noyz Narcos

Abstract

Social media are pervasive in our daily lives. These platforms are gaining momentum thanks to powerful artificial intelligence models. Serious negative consequences of social media have been repeatedly highlighted in recent years, pointing at various threats to society and its more vulnerable members, such as teenagers, in particular, ranging from much-discussed problems such as digital addiction and polarization to manipulative influences of algorithms and further to more teenager-specific issues.

The thesis proposes a collective well-being-oriented social media companion by examining the background and context surrounding these issues, such as problems related to the AI-human value alignment and how to deploy Educationally managed social media and well-being metrics that are flexible enough to take into account specific needs and the coexistence of different stakeholders such as educators and content creators.

The research includes agent-based simulations to examine the impact of content and people recommendation strategies on opinion dynamics and otherwise invisible network-based threats directly into the user feed, such as polarisation and echo chambers.

A deep learning architecture based on a graph neural network for classifying tweets is introduced, and a dataset annotation protocol that combines objective and subjective features to understand factors affecting the potential misleadingness of images is proposed because methods for addressing social media threats must support multimodal content too.

Overall, the results indicate that machine learning models can serve as detectors for labelling content in the proposed companion and that there is room left for including additional types of information (e.g. social media context) given the flexibility of the proposed architecture. The game-based digital media literacy educational intervention inspired by the *wisdom of crowds* phenomenon can increase the perception of social media's influence on participants, so it fits educational needs.

This contribution is motivated by the desire to improve the impact of social media on society. The presence of a trade-off between users' rights and duties or freedom VS safety introduces ethical issues that require the formulation of a comprehensive and shared view of the values of the social media community, so more multidisciplinary work is still required.

Keywords: social media, recommender system, teenager, echo chamber, filter bubbles, educational activities, deep learning

Contents

List of Figures	vii
List of Tables	x
1 Introduction	1
1.1 Motivation	5
1.2 Problem Statement	7
1.3 Thesis Outline	8
1.3.1 Publication List	9
2 Background and Preliminaries	11
2.1 Social media are the modern Janus Bifrons	11
2.1.1 Difference between social media and social network	12
2.2 Introduction to Machine Learning: Fundamental Concepts and Techniques	13
2.2.1 The basic building block of ML: the data	14
2.2.2 From data to Dataset	15
2.2.3 The problems	17
2.2.4 How to evaluate: Loss functions	18
2.2.5 How to learn: Optimizer	19
2.3 Overview on Recommender Systems	22
2.3.1 Tasks and Evaluation Metrics for Recommender Systems	23
2.3.2 Collaborative filtering	25
2.3.3 Content-Based	26
2.3.4 Hybrid recommenders	27
2.3.5 Neural Network Recommenders	27
2.4 Conclusion	28
3 Navigating Social Media Risks and Recommender System Biases: A Case for Human-Centered AI	29
3.1 Social Media Threats classification	29
3.1.1 Individuals: cognitive limits and emotional state	30
3.1.2 Harmful content characterisation	31
3.1.3 A solution to harmful content: Threat Detectors and Content Analyzers	33
3.1.4 Text-Based Detectors	33
3.1.5 Auditing Algorithms	34
3.2 Recommendation and potential biases	38
3.2.1 Freedom to speech and freedom to reach	40
3.2.2 Information personalisation algorithms and transparency issues	40
3.3 Toward a more transparent information environment	42
3.3.1 Are social media utilities?	42
3.3.2 Recommender systems and their role as editorial board	43
3.3.3 Improve models: going beyond accuracy	44

3.4	AI and human value alignment	45
3.5	Alignment and Definition of Collective Well-Being Values	47
3.6	Educational Social Media Companion	49
3.7	An Educationally Managed Social Media Community	50
3.8	Defining a Collective Well-Being Metric for Social Media	53
3.8.1	Research on collective well-being and social media	53
3.8.2	Participative definition of social media community principles and CWB factors	55
3.8.3	Toward the automatic estimation of collective well-being in social media communities	56
3.9	An educational Collective Well-Being Recommender System	58
3.10	Conclusion	59
4	Exploring the Interplay between Social Media and Recommender Systems through Simulation	60
4.1	Agent-based models as Mathematical Representations of dynamic society	60
4.1.1	Advantages of Agent-based models	61
4.1.2	Disadvantages of Agent-based models	61
4.2	Agents and Opinion Spaces	62
4.2.1	Binary Opinion	62
4.2.2	Continuous 1D Opinion	62
4.2.3	Continuous 2D Opinion	62
4.3	Society as a graph	63
4.3.1	Graph theory elements needed for Agent-based models	63
4.4	Modelling Interactions	64
4.4.1	DeGroot Model	64
4.4.2	Bounded confidence model	65
4.4.3	The Friedkin–Johnsen model	66
4.5	Recommender systems in simulation settings	67
4.6	Proposed Approach	67
4.6.1	Network Generation	68
4.6.2	Recommendation strategies	69
4.6.3	Simulation protocol	70
4.7	Results, limitation and future works	71
4.7.1	Recommending people	72
4.7.2	Content Filtering	74
4.8	Apply real data to ABM simulations	75
4.9	Conclusion	76
4.9.1	Limitation and future works	77
5	Digital Media Literacy and Information Personalization: A Game-Based Approach for Teenagers	78
5.1	Information personalisation and decision making	78
5.1.1	Digital Media literacy and remote teaching impact	79
5.2	Wisdom of folly of the crowd?	79
5.3	Related works	80
5.3.1	Digital Media Literacy and Teenagers	80
5.3.2	Recommender Systems and Social Media	80
5.3.3	Wisdom of Crowds	81
5.4	Experimental Methodology	82
5.4.1	Participants	83
5.4.2	Digital Media Literacy Talk	83
5.4.3	Wisdom of crowds game experience	84

5.5	Results and discussion	85
5.6	Conclusion	88
5.7	Limitation and future works	88
6	A Deep Learning Approach to Identifying Harmful Content using Graph and Word Embeddings	89
6.1	Introuction	89
6.2	Related work	90
6.3	Proposed model	91
6.3.1	Graph creation	92
6.3.2	Node characterisation	92
6.3.3	Graph attention convolution and max pooling layer	93
6.3.4	Dense	93
6.4	Experimental setup	93
6.4.1	Dataset	93
6.4.2	Model training	94
6.5	Results	94
6.5.1	Baseline	94
6.5.2	Proposed Approach	94
6.6	Discussion and future work	95
6.7	Conclusion	95
7	Misinformation through images	96
7.1	Introduction	96
7.2	Motivation	97
7.2.1	The role of computer-generated content	97
7.2.2	Collaboration between publisher and social media to counteract visual misinformation	98
7.3	Why true false dichotomy is not sufficient	98
7.4	Factors that contribute to images misinterpretation	99
7.4.1	Background knowledge and skills	99
7.4.2	Images manipulation	100
7.4.3	Subjective factors that influence inteprretation	102
7.4.4	Staged events	102
7.4.5	3D reconstruction and beliefs misinterpretation contribute to misleadingness .	103
7.5	A richer characterisation of image features	103
7.5.1	Authenticity	103
7.5.2	Truthfulness	106
7.5.3	Declared Manipulation	107
7.5.4	Detected Manipulation	108
7.6	A crowd-sourced subjective characterisation of images	108
7.6.1	Authenticatablity	109
7.6.2	Credibility	109
7.6.3	Manipulation Visibility	110
7.7	Misleadingness	110
7.7.1	Uninformed Misleadingness	112
7.7.2	Informed Misleadingness	113
7.8	Dataset Annotation procedure	113
7.8.1	Dataset Description	114
7.9	Annotation results	115
7.10	Conclusion	117
7.10.1	Future work	117

8 Conclusions

118

Bibliography

120

List of Figures

1.1	Number of social media users worldwide from 2018 to 2027 (in billions). Source: statista.com	2
1.2	<i>Left</i> : traditional <i>Unidirectional</i> media information flow. <i>Right</i> : new media information flow where everyone can directly expose or be exposed to information from connected peers across different platforms.	3
1.3	Multi-Sided Platforms structure: different classes of users, characterized by diverse objectives and needs, are served by the platforms. A typical example is LinkedIn, the world’s leading professional networking service, which currently runs a three-sided platform that connects individual users (professionals), recruiters and advertisers. . .	4
1.4	Each node (dot) is a member of the parliament. Colours represent political affiliation, and each edge is drawn if nodes agree above the Congress’ threshold value of votes. Source: Andris <i>et al.</i> , [25]	10
2.1	Florentine families network during the first half of the 14 th century. Colours indicate the Eigencentrality value developed by [20]	13
2.2	In the picture, a representation of a social network as an attributed and directed graph is depicted	16
3.1	RS main component: Users, Data, Model, and the representation feedback-loop mechanism. Once the model is learned, users provide feedback that is used to refine the model.	39
3.2	<i>Sketch of Companion User Interface</i> The Companion will support the students’ interaction with social media by contextualizing the content to increase students’ awareness and allow them to access a more diverse set of perspectives [58] and sources. It also explicitly and visually provides the students with an evaluation of the content’s harmfulness [135]. The example shows how a piece of imaginary fake news would be contextualized.	51
3.3	<i>Role of the CWB-RS in the Companion.</i> CWB-RS will process the <i>content generated by the users</i> of the <i>educationally managed social media</i> and the <i>content externally recommended</i> for them by the RSs of the external social media platform to create new recommendations aimed at maximizing the cumulative long-term <i>collective well-being metric</i> . <i>Content Analyzers and Threat Detectors</i> will analyze and evaluate the level of threat for each piece of content and other relevant information as the users’ emotional state. This information will be used to (1) <i>augment</i> the information provided to the users by the companion interface; (2) <i>evaluate</i> through <i>predictive models of users’ opinions and reactions</i> the future effects of different sequences of re-ranking and recommending actions; (3) <i>select</i> the re-ranking and recommending actions that resulted in the highest expected cumulative improvement in terms of learning objectives, CWB metrics, agreement with selected educational strategies and user engagement.	59

4.1	Opinion spaces: binary opinion (top right), Binary continuous opinion (top left), 2D continuous opinion (bottom right), 3D opinion (bottom left).	63
4.2	Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.	72
4.3	Feed Satisfaction is computed as a function that weights close and distant posts based on the β value.	73
4.4	Polarisation tends to increase for all the strategies.	73
4.5	Disagreement is a local metric computed in each neighborhood	73
4.6	Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.	73
4.7	Feed Satisfaction.	74
4.8	Polarisation	74
4.9	Disagreement	74
4.10	Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.	75
4.11	Feed Satisfaction represents the utility of the feed.	75
4.12	Polarisation.	75
4.13	Feed Entropy	75
5.1	The proposed experimental protocol. The total expected time for completion of the activity is 1h and 30 minutes.	83
5.2	Sample of the slides presented to participants.	85
5.3	Structure of the game-based wisdom of crowds estimation task. It starts with the question where the image on the top right is shown and they are asked to estimate the number of dots in it. After the aggregate of individual answers, as social information, is disclosed to the participants they can provide the final answer.	86
5.4	The blue histogram reports the PRE-activity questionnaire, while the orange one the POST-activity	86
6.1	Model parameters (Top) numbers in brackets indicate parameters' tensor dimensions; the last column indicates the number of parameters in each layer. Model architecture (bottom) model input and output shapes in each layer (figure taken from the Google Colab notebook).	91
6.2	Graph representation: each tweet is represented as a graph after pre-processing and POS tagging	92
7.1	An image of Superman and Ironman fighting in the sky above New York generated by the AI model DALLE-2.	99
7.2	Christopher Reeve playing Superman flying taken from the official movie	100
7.3	The portrait of Dora Maar is a 1937 oil on canvas painting by Pablo Picasso. It depicts Dora Maar, the painter's lover, seated on a chair.	101
7.4	A woman portrait in Picasso style generated by the AI DALLE-2	101
7.5	A GAN-generated image of a Politician	101
7.6	An image taken from the movie 1917, directed by Sam Mendes, that could be wrongly interpreted as a real photograph taken during a battle of WWI or WW2.	103
7.7	The Sun Cruise Resort&Yacht, located in Jeongdongjin on the east coast of South Korea, photographed from the beach.	104
7.8	The famous 'Hitler moustache' that casts a shadow over Merkel-Netanyahu meeting on February 25, 2014	111

7.9	Flowchart of the dataset annotation process.	113
7.10	A screen of the Qualtrics annotators page.	114
7.11	The table reports the correlation between images features	116
7.12	Polar graph of truthful (Left) and not truthful images (Right). Each pole represents an annotated feature.	116

List of Tables

- 1.1 The Table reports some examples of *Web1.0* services and their *Web2.0* counterpart. 3
- 4.1 The table reports the metrics used to evaluate the model’s output 72
- 4.2 Each row of the table is a dataset, and different features are reported in the columns. 76
- 5.1 Perceived Social Media Influence before and after the intervention. Columns represent the two items: (I) Influence on them and (ii) Influence on their peers. **P-Values** are reported and * means the significance level at 0.005 87
- 5.2 Perceived Social Media Influence before and after the intervention. Columns Self and Other represent the target of the question related to the perceived influence of social media on themselves or their peers. Values in cells correspond to the average, while the sample size is reported in brackets. 87
- 5.3 The table reports the contingency tables for each item proposed (Perceived Influence on themselves and Perceived Influence on peers). In each row, the number of participants with an increasing (decreasing or not answered) difference between the initial and final survey is reported for each condition (baseline and experimental). 87
- 5.4 The table reports the p-values for each item in the survey (rows). Columns report both cases where unanswered questions are either considered a decrease or a class. 88
- 6.1 Dataset statistics of all provided splits for English. 94
- 6.2 Results (binary Precision, Recall and F1 of the positive class label) on the official test set for English with respect to different approaches. 95
- 7.1 Table reports in each row an *Objective or Subjective* feature, and the third column provides a description of the feature. 104
- 7.2 The table reports the sample size for each topic and is also divided between truthfulness and authenticity. 115
- 7.3 The table reports uninformed and informed misleadingness for images divided into Above (*High*) and below (*Low*) the average credibility while annotators are divided into above (*High*) and below (*Low*) the average critical thinking. 117

Chapter 1

Introduction

Digital platforms based on Web 2.0 [111], an umbrella term that refers to websites that emphasize **user-generated content** ¹ such as social media, nowadays dominate the information and entertainment arena all around the world. Projections to 2027 ² reported in Figure 1.1 clearly show an increase in the share of active users on these platforms.

Understanding the outcome of this collective participation and its effects on individual well-being and welfare is crucial to derive consensus and extensive adoption of a policy framework. A disruptive regulatory innovation [294] will be required in the following years to safeguard society and exploit the benefits and positive aspects of social media for citizens, institutions and private companies.

The first theme to consider in this regulatory framework is deeply intertwined with the business model these platforms introduced. The so-called phenomenon of platformisation refers to the increasing centralization of digital markets around a few dominant platforms. In Figure 1.3 the actors involved in multi-sided platforms are depicted. Social media platforms like Facebook, Twitter, and Instagram have become dominant players in the digital advertising market, with Google and Meta controlling over half of the global digital advertising revenue in 2019. This concentration of market power has led to concerns about the ability of these platforms to control access to information, and undermine competition.

The dominance of a few large platforms has led to concerns about the lack of competition, with smaller players struggling to compete and innovate in the digital advertising market. This concentration of market power has also led to concerns about the ability of these platforms to engage in anticompetitive practices, such as using their dominance to leverage into adjacent markets or acquire potential rivals.

This aspect of the regulation must address these concerns by creating a level playing field for competition, preventing anticompetitive practices, and promoting innovation. This could involve measures such as imposing data interoperability requirements, promoting open standards, and mandating platform neutrality in their fee structure.

The second theme to consider is the role of social media as a public good. Social media platforms have become important tools for communication, information sharing, and social interaction, with many users relying on these platforms for access to news, public services, and social connections.

¹A definition of web2.0 can be found here https://en.wikipedia.org/wiki/Web_2.0.

²Source: Statista.com <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>.

However, the dominance of a few large platforms has led to concerns about the privatization of public discourse and the lack of democratic accountability given also the recent diffusion of laws that impose practices of business and algorithmic auditing.

A disruptive regulatory framework is needed to address these concerns by recognizing the public value of social media and promoting public interest objectives. This could involve measures such as mandating platforms to disclose their algorithms and decision-making processes, promoting user data portability, and establishing public oversight mechanisms. At the moment, only Twitter open-source its content Recommender systems³. Moreover, the balance between public and private interest should be balanced based on democratic decisions that must involve all the stakeholders.

In addition to the regulatory framework proposed above, it is worth exploring alternative models for social media that prioritize social management and learning. The proposal of this thesis will go in the direction of creating an educationally managed social media community where educators, users, and all stakeholders have a voice and can actively transform social media into a learning environment. By creating a social media that prioritizes education and community-building, it is possible to harness the potential of social media to promote critical thinking, civic engagement, and digital literacy. This proposal aligns with the public value proposition of social media as a tool for social connection and collective action, while also addressing concerns about privacy, competition, and democratic accountability.

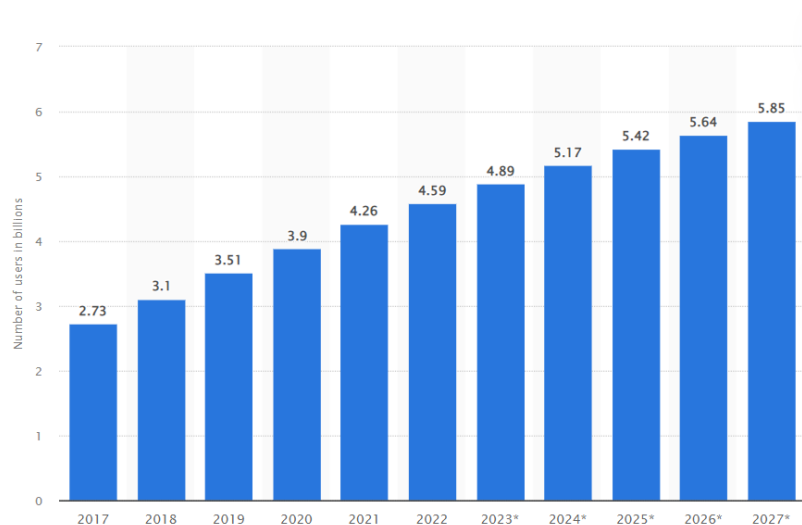


Figure 1.1: Number of social media users worldwide from 2018 to 2027 (in billions). Source: statista.com

The shift in the communication paradigm, which characterizes social media platforms, described by Baeza-Yates *et al.*, [29], highlighted that people are theoretically free to connect directly with others in this new era. They could publish whatever they wanted without intermediaries between themselves and their (connected) peers as early as 1999. *Traditional* (such as radio, television and newspaper) and *new media* (Digital platforms) information flows are depicted in Figure 1.2. It points out the role of each user in his context, which can dramatically influence what information he or his connected peers are exposed to.

This participatory aspect and the emphasis on user-generated content are critical components

³Check here: https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm, retrieved on April 7, 2023

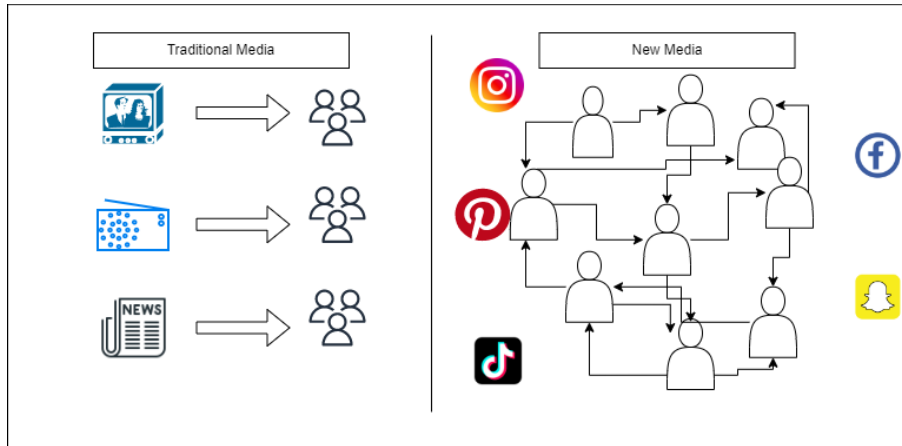


Figure 1.2: *Left:* traditional *Unidirectional* media information flow. *Right:* new media information flow where everyone can directly expose or be exposed to information from connected peers across different platforms.

of social media platforms. People can connect with each other by creating a network based on relationships and interests (collaborative or competitive). They can exchange multi-modal (including text, images, and videos) information and user-generated content while giving explicit or implicit feedback through likes, shares, clicks, and ratings. These characteristics make each user a self-publisher, and these new platforms' paradigms, also called "*network as platforms*", substitutes the equivalent Web 1.0 on the web. In Table 1.1, some Web1.0 and Web2.0 counterpart examples are reported.

Web1.0	Web2.0
mp3	Napster
Britannica Online	Wikipedia
Ofoto	Flickr
personal websites	Blogging

Table 1.1: The Table reports some examples of *Web1.0* services and their *Web2.0* counterpart.

From an economic perspective, digital platforms are characterized by lower search and transaction costs⁴ [123], which has allowed them to become a leading business model. The platforms' users (who can be divided into one or more groups, called sides) are affected by direct and indirect network effects, which means that users' experience (and their utility in economics terminology) can be affected simply by the number of participants. Moreover, the fee structure (eventually) applied to each side can be non-neutral. This perspective highlights the relevance of taking a comprehensive approach when evaluating the impact of multi-sided platforms and their algorithm engines, such as social media. Wang *et al.*, [424], fund at least two challenges for multi-sided platforms concerning recommender systems:

1. Different, and potentially conflicting, utility functions for each side
2. in the case of heterogeneous item space, the classical recommender systems algorithms can

⁴Social media can reduce the transaction costs to exchange information with friends because they can reduce information asymmetry while platforms like eBay or Amazon reduce the search costs, making it easier for buyers and sellers to find each other.

fail because recommended items are a collection of other recommendations (i.e. the current Netflix recommender system)

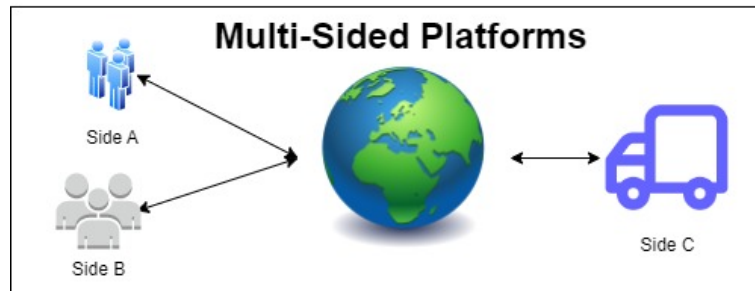


Figure 1.3: Multi-Sided Platforms structure: different classes of users, characterized by diverse objectives and needs, are served by the platforms. A typical example is LinkedIn, the world's leading professional networking service, which currently runs a three-sided platform that connects individual users (professionals), recruiters and advertisers.

The increasing rate of use of social media that, in most cases, is free to use was driven by the lowering cost of multiple factors:

- Internet access
- Mobile devices
- Mining, storing, processing and analyzing data

The widespread adoption of smartphones allows everyone to overcome every geographical limit and potentially interact with anyone with a smartphone and a connection. The so-called "big data" is the Internet's, mainly social media's, most exciting by-product. A simple definition, based on Mills *et al.*, [266] is:

"[Big Data is] a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information".

The colossal volume (in terms of quantity of data), low veracity [189, 345] (intended as the uncertainty included in the data), and also the low density of information underlying the data all set tremendous challenges to developing automated technologies to process, analyze and support strategic decisions efficiently. Still, they also create significant opportunities to generate profits for those who control all the process phases.

This definition, and the current debate on the need for transparency ⁵, point out the leading role of tech companies and urge a further understanding of social media's overall effect on people's well-being.

There is no doubt that there are many positive effects of social media. However, in recent years, some of its seriously negative implications have repeatedly been highlighted, such as its potential threat to society and, more specifically, to its most vulnerable members, such as teenagers. These

⁵See for example the Wall Street Journal investigation "Facebook Files" here <https://www.wsj.com/articles/the-facebook-files-11631713039>, retrieved on April 7, 2023

potential drawbacks range from digital addiction [18] and polarization [146] to the manipulative influences of algorithms and teenager-specific issues (e.g. body stereotyping) [251].

A recent investigation by the Financial Times ⁶ can be helpful to summarize what is already clear: everywhere you turn, adolescent mental health is in decline, and the turning point—give or take a year or two—was in 2010, when smartphones went from being a luxury to being commonplace. In [149] find that Depressive Affect (DA) scores rose after 2010, but increases in female Liberal Adolescents were the most pronounced. Correlation is not causation, however, a growing body of studies [15, 16] indicates that spending less time on social media is good for mental health.

User activity and the filtering process operated by intelligent platform components such as recommender systems determine social media’s functioning and, therefore, the impact on both the individual and social levels. Thus, users must understand both the role they and other users play in these cybercommunities and social media algorithmic mechanisms. On the other hand, even with increasing educational efforts on digital citizenship in schools and related contexts, it is not realistic to expect that all users will reach a level of empathy and media literacy that will allow for the disappearance of harmful content and behaviours from social media platforms. Even with a high majority of educated users, it is essential to consider that recommender systems can distort the perceived community structure in any domain, [48, 91] giving high visibility to toxic behaviour, and that the overabundance of information delivered by social media can overload users’ cognitive limits [337].

Finally, social media platforms should also favour and invest in resources to develop and deploy machine learning algorithms that can manage the so-called *AI and human value alignment problem*, which refers to the challenge of ensuring that the goals and values of artificial intelligence systems are aligned with humans ones [136]. This issue is particularly relevant when it comes to recommender systems [381], which are potent engines of information spreading that may reinforce existing biases and perpetuate harmful stereotypes, leading to a lack of representation and the marginalization of certain groups, thus promoting further polarization of opinions and the creation of echo chambers. Therefore, if the platforms are not aimed at social good and the users’ well-being but at arbitrary self-referential objectives (profit or profit-related metrics), they may negatively impact society by overflowing users with toxic but engaging content released by malicious or uneducated actors and expose users to severe and unnecessary emotional conditions, such as flames and cyberbullying, against which education may not be enough.

1.1 Motivation

The advent of social media and the consequent abundance of data allows companies to train machine learning-based models ⁷, (see 2.2 for a deeper discussion of the topic.) that learn based on billions of examples and can provide personalized "information diets autonomously" based on user interests and preferences to help surf the tremendous amount of news, photos, blogs and videos uploaded each minute to find the required information and content or at the least, an item that will grasp

⁶Check Here: <https://www.ft.com/content/0e2f6f8e-bb03-4fa7-8864-f48f576167d2> retrieve on April 7, 2023

⁷The term "Machine Learning" indicates a family of algorithms where machines indicate an algorithm that can process examples of input-output couples and estimate parameters solely based on a list of examples reaching the capabilities to generalize (the "Learning") a set of "rules" expressed through the model’s parameters that can infer unseen examples after training.

user’s attention and convert it into an interaction that can enrich user profile models or be decorated with personalized advertising.

Social media core business relies on these algorithms that allow for auctioning off personalized ads (tailored to users’ profiles to maximize their attention and likelihood of buying). Still, discriminatory outcomes are present even in ads intended to be gender-neutral [225]. Threats can take shape on different levels, such as content (fake news, hate-speech), dynamics factors such as an increasingly polarized society and algorithmically induced ones that stem from the fact that these algorithms try to maximise what is profitable for the company that trains them and this sometimes differs from what would be considered optimal for a community ⁸. It is unclear how to balance the evident effectiveness of artificial intelligence (AI) algorithms in specific tasks with a more comprehensive and long-term-oriented evaluation strategy. This strategy should also consider the user’s current state, cognitive limits and style and social and information context while trying to orient his or her experience towards learning and well-being rather than the repeated consumption of items or relationships. The relevance of the sociopolitical consequences of the impact of social media increasingly clashes with the public interest: political actors have begun to question tech giants since the Cambridge Analytica scandal that forever changed the sensitivity of public opinion ⁹ [191].

More specifically, on the user side, it is crucial to understand which factors are taken into account by machine learning models. In particular, information filtering systems, because models could prioritize explicit and deliberate feedback where there is a high degree of awareness of the user, i.e. actions like *following* rather than unconscious factors (such as video view rate), will put users more in control and give them the possibility to understand which types of feed-back influences the system and how to interact with it actively.

Alongside the impact on society, there are also multiple key business issues: the credibility and perception of social media companies and, more specifically, the algorithmic-driven consumption that could be inefficient for both companies and users or other companies that buy services on the platform. A system that recommends songs could reduce the diversity of content users listen to, even if it increases short-term business metrics. On the contrary, it is interesting to point out that crucial long-term metrics positively correlate with a diversified listening track record, as highlighted by Anderson *et. al* [23].

Moreover, it is unfair to blame social media for contemporary society’s difficulties and challenges. There have long since been clear indications that society was moving toward increasing partisanship. What people see in social media is probably only the digital counterpart of a broader process.

Figure 1.4, taken from [25], depicts the increasing partisanship in the United States parliament, expressed through the number of times that representatives gave the same vote, thus highlighting the already clear signs that society is becoming more divisive.

To shape the influence of social media on society, and especially on teenagers, an educational

⁸Seattle Public Schools (SPS) sues social media firms over youth mental health crisis asserting the companies are substantially contributing to a youth mental health crisis. The 90-page lawsuit, filed in U.S. District Court in Seattle, alleges the social media companies intentionally market, design and operate their platforms to maximize engagement from young users for profit. Check here <https://www.seattletimes.com/education-lab/seattle-schools-sues-social-media-firms-over-youth-mental-health-crisis/>, retrieved on April 7, 2023

⁹See the transcript of Meta Platforms, Inc. CEO, Mark Zuckerberg’s audition at the Senate’s Commerce and Judiciary committees here, (retrieved on April 7, 2023) <https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/>.

approach should be adopted to promote a safer and healthier digital environment characterized by educationally and algorithmically defined learning strategies that can go beyond accuracy, including AI, and also foster diversity and novelty in what and who is recommended to users [199]. In conclusion, machine learning algorithms that are more aligned with human values can be beneficial also for social media platforms because they can build trust and credibility with users. When users understand how and why an AI system prompts a particular recommendation or action, they are more likely to trust and accept it. This can lead to higher levels of engagement and satisfaction, which can benefit the platform too.

1.2 Problem Statement

Digital platforms, mainly social media such as Facebook, Twitter, Instagram, and TikTok, are pervasive in people’s daily life, allowing users to connect, share information and content and participate in online communities. At the same time, they are becoming the primary source of information and entertainment tools used to maintain social networks and a key business stakeholder in all developed and developing countries.

Social media are centred on user-generated content and interests. Machine learning models are the main drivers of information diffusion: given the overwhelming amount of news, interactions and items that are present online, the best pieces of information must be found, filtered, and ranked by models to distil only the information that is considered **relevant** for each user in a personalized manner. News (and content in general) and other users can be added to the ego network, providing more opportunities to interact.

Recommender systems are algorithms that aim to predict users’ interests and provide personalized recommendations based on their past behaviour and preferences. These systems are used by various online platforms, including e-commerce sites, streaming services, and news websites, to help users discover new products, movies or articles they may be interested in. Recommender systems have become essential for businesses looking to increase customer engagement and sales. They can help surface relevant content and products that are more likely to interest-specific users. Such algorithms can affect society in multiple and complex ways, from bias in the training data to the interplay with cognitive and dynamic factors.

This thesis addresses the problem of modelling social media users and the threats they can be exposed to on multiple levels with the goal of raising users’ awareness. Solutions that allow for creating educationally managed social media communities have been developed through attempts to identify the algorithms and dynamic components that can affect the collective well-being on social media.

The goal is to implement these models into a virtual social media companion for teenagers developed inside the COURAGE project, which is a multidisciplinary consortium ¹⁰.

Simulations were used when data availability was limited, and experiments were conducted with real students. In particular, this thesis tries to answer the following questions:

- Which elements compose a social media companion for teenagers oriented toward collective well-being?

¹⁰Source here: <https://portal.volkswagenstiftung.de/search/projectDetails.do?siteLanguage=en&ref=95563>

- What is the impact of directly experiencing the algorithm’s influence (in the decision-making process) inside a digital literacy activity concerning the perceived influence of social media on teenagers?
- How to model complex network and opinion dynamics that mimic specific social media features, such as backfire and recommender systems, in a simulation environment?
- How to characterise misleading images without context while taking advantage of fact-checked and annotated features?

1.3 Thesis Outline

This thesis examines social media threats that can arise from users, content and artificial intelligence algorithms, focusing on mitigating biases and developing strategies and activities to reduce the impact of threats and raise teenagers’ awareness. The thesis is composed of both theoretical (2, 3) and empirical chapters (4, 5, 6,7).

The **introduction** has provided motivations, framed the research gap, and presented this work’s structure and publication list.

The **background and preliminaries** chapter provides the needed preface and background knowledge along with the basics of machine learning (presenting the most common problem set and algorithms used to allow neural networks to learn), introducing the most common techniques to provide recommendations (based on the interaction between users and items or on content similarities).

The **third chapter** analyses users’ biases and potential threats of social media content and recommender systems. Here is introduced the concept of collective well-being-oriented social media inside a social media companion framework. The problems that should be considered to develop a metric suited for this goal, concerning the AI-human value alignment problem and the presence of multiple stakeholders that characterize the social media environment, are addressed.

The **fourth chapter** uses agent-based modelling to simulate social network dynamics and the impact of different people and content recommender strategies in favouring the creation of echo chambers and polarisation for a given opinion model. These experiments allow studying emerging phenomena while controlling for multiple parameters which govern agents and recommenders.

The **fifth chapter** presents an interactive game-based digital media literacy activity designed to raise student awareness of social media threats. The task here involved designing and measuring the effectiveness of the educational intervention.

The **sixth chapter** explores the application of deep learning architecture to detect harmful tweets. Here a model based on Graph Neural Networks and word embedding is used in a binary classification task.

Finally, the **seventh chapter** conceptualizes objective and subjective features applied to images with low or absent context, specifically designed for the context of visual misinformation. The proposed conceptualisation has been tested in a pilot study with a sample of annotators. The outcome of this process is an annotated dataset that can be used to perform different activities that can help to understand users’ susceptibility to visual misinformation.

1.3.1 Publication List

Publications that contribute to this thesis:

- D. Ognibene, R. Wilkens, D. Taibi, D. Hernández-Leo, U. Kruschwitz, G. Donabauer, E. Theophilou, F. Lomonaco, S. Bursic, R. A. Lobo, and et al. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence*, Dec 2022
- F. Lomonaco, D. Taibi, V. Trianni, and D. Ognibene. A game-based educational experience to increase awareness about the threats of social media filter bubbles and echo chambers inspired by “wisdom of the crowd”: preliminary results. In *Book of Abstracts*, page 84, 2022
- F. Lomonaco, G. Donabauer, M. Siino, et al. Courage at checkthat! 2022: Harmful tweet detection using graph neural networks and electra. *Working Notes of CLEF*, 1, 2022
- T. Emily, S. Veronica, B. Johanna, S.-R. J. Roberto, S. Lidia, L. Francesco, A. Farbod, O. Dimitri, T. Davide, H.-L. Davinia, and S. Eimler. Empirically investigating virtual learning companions to enhance social media literacy. *Fulantelli et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022.*, 2022 (**Accepted, to appear in March**)
- A. Farbod, M. Nils, L. Francesco, D. Gregor, O. Dimitri, K. Udo, H.-L. Davinia, F. Giovanni, and H. H. Ulrich. The “courage companion” - an ai-supported environment for training teenagers in handling social media critically and responsibly. *Fulantelli et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022.*, 2022 (**Accepted, to appear in March**)
- D. gnibene, G. Donabauer, U. Kruschwitz, R. S. Wilkens, S. Bursic, D. Hernandez-Leo, E. Theophilou, F. Lomonaco, and U. Kruschwitz. Moving beyond benchmarks and competitions: Towards addressing social media challenges in an educational context. *Datenbank-Spektrum*, 22(1):5–15, 2023 (**Accepted, to appear in March**)

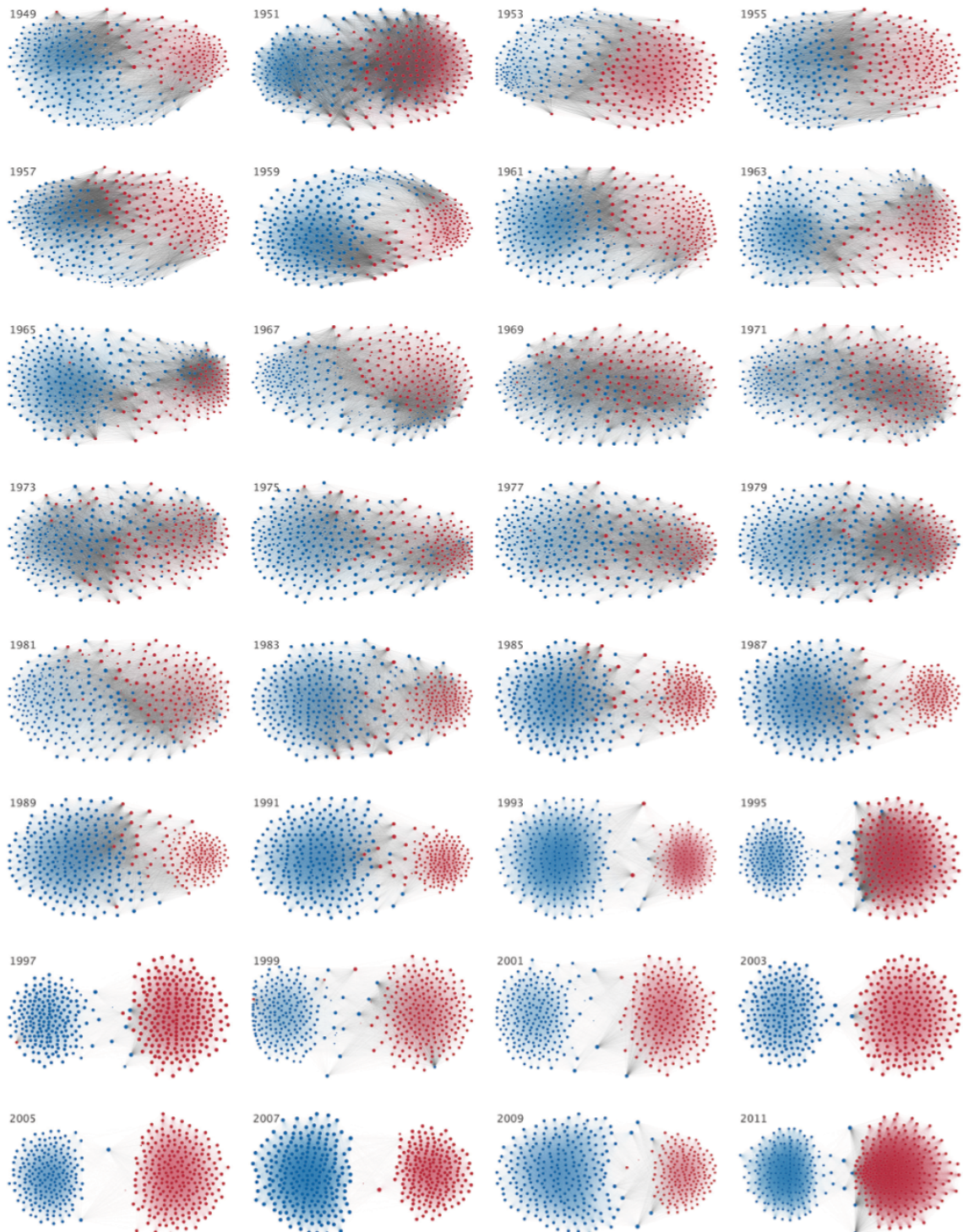


Figure 1.4: Each node (dot) is a member of the parliament. Colours represent political affiliation, and each edge is drawn if nodes agree above the Congress' threshold value of votes. Source: Andris *et al.*, [25]

Chapter 2

Background and Preliminaries

In this chapter, a brief preface and background knowledge on the topics of machine learning (ML) and recommender systems (RS) are introduced. Machine learning is defined, and its main categories are discussed, including supervised, unsupervised, and reinforcement learning. The basics of recommender systems are then introduced, including the types of recommendation algorithms and the challenges and limitations of these systems. The process of training and evaluating machine learning models for various tasks is discussed. Overall, this chapter serves as a foundation for understanding the basics of machine learning and recommender systems, which will be explored in greater depth in subsequent chapters.

2.1 Social media are the modern Janus Bifrons

Over the past few years, social media marked a new beginning and have become the primary means of information and social relationships. Drastic changes opened many new possibilities and opened the space to pitfalls for users. Like the ancient roman god *Janus Bifrons*, the god of beginnings, and transitions, that frequently symbolised changes such as the progress of past to future, social media too could be represented by two opposite faces. One side tells us that social media can benefit users in multiple ways. For example, improved relationship maintenance [115], reduced loneliness [207, 348] but also giving access to tons of open access libraries, educational material, and news broadcasters at an unseen low price. Moreover, digital platforms enable the aggregation and the creation of communities of activists [35, 184] like during the Arab Spring in the early 2010s or the more recent #TulsaFlop [35] during 2020 US presidential elections. Given the extensive adoption by institutions and politicians, social media platforms are also the easiest way to access the public sphere at a low cost because only a smartphone with internet access is needed.

Alongside these positive aspects and despite social media, companies are trying to enforce regulations to avoid drawbacks and re-align the long-term objectives with society taking initiatives such as the adoption of IFCN (International Fact-Checking Network) certification for outsourced fact-checkers provided by the Poynter Institute based on the application of the code of principles ¹.

The other face points out the urgency to have a comprehensive understanding of the effects of social media on society in the long term, especially for politics-related issues such as radicalisation.

¹Read more about the Fact-checkers certification required for external sources which collaborate with Google and Meta Platforms here <https://www.ifcncodeofprinciples.poynter.org/know-more>.

The duality of *Janus Bifrons*, the guardian of transitions, reflects that social media platforms can be seen as facilitators of change, connecting us to both the past and the future. However, it is essential to acknowledge that social media platforms can also bring with them shadows and grey areas, blurring the lines between the past and the future and raising questions about privacy, identity, and the impact of technology on society.

Even if polarisation, a situation where hostile feeling for the opposite opinion or condition (i.e. political leaning or race) [192], in contemporary society is a matter of fact and could also be found in data (see figure 1.4), it is not clear if a recommender system (RS) can boost user radicalisation for example on YouTube [182, 446]. An attempt to include also a depolarising aim in the recommender has been proposed by Stray *et al.* [380], but more efforts from different fields are needed.

Moreover, creators too started asking to raise awareness of users concerning biases that can affect them [306] but also to become a recognised profession by State authorities ².

2.1.1 Difference between social media and social network

Social media and social networks are often used interchangeably, but they are two different things. Social media refers to the platforms and technologies that enable users to create and share content or participate in social networking online. Examples of social media include Facebook, Twitter, YouTube, TikTok and Instagram.

On the other hand, a social network is a network of individuals connected by certain social relationships. These relationships can be based on various factors, including common (or competing) interests, shared values, or personal connections. Social networks can exist online and offline and can be small and tightly knit or large and sprawling. An example of a famous social network is the so-called *Florentine Family network* of the 14th century. The social relationship was based on marriages (the easiest and most effective way to establish strong ties between families that ruled the city in that period), allowing the De Medici family to become the town's leader. Figure 2.1 below reports the connections between families elaborated by [20], and the data are taken from [302].

One key difference between social media and social networks is that social media is a tool that can be used to facilitate the formation and maintenance of social networks. In contrast, social networks are the actual connections and relationships between individuals. In other words, social media is a means to an end, whereas social networks are the end itself. Another key difference is that social media is often associated with the sharing of user-generated content, such as photos, videos, and comments, whereas social networks are more focused on the connections and relationships between individuals. While social media and social networks are closely related, they are distinct concepts with unique features and characteristics. Social media is a tool that can be used to facilitate the formation of social networks, while social networks are the actual connections and relationships between individuals.

²In Italy, the House of Representatives discussed inside the Labor XI Permanent Commission to provide a first contribution in terms of proposals for the strengthening of the protection framework for operators in the sector. Check here the transcription of the commission audition of Andrea Panciroli, Ivan Grieco, e Sara Stefanizzi (creators) http://documenti.camera.it/leg18/resoconti/commissioni/stenografici/pdf/11/indag/c11_contenuti_digitali/2021/05/06/leg.18.stencomm.data20210506.U1.com11.indag.c11_contenuti_digitali.0002.pdf, retrieved on April 7, 2023

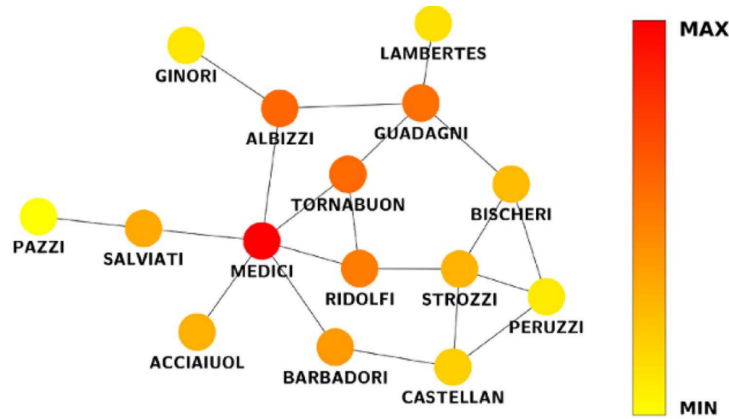


Figure 2.1: Florentine families network during the first half of the 14th century. Colours indicate the Eigencentrality value developed by [20]

2.2 Introduction to Machine Learning: Fundamental Concepts and Techniques

As previously highlighted, the focus of social media is on user-generated content. The amount of information produced by users must be classified and labelled to fit the input requirements of machine learning models. Given the vast amount of information and the capability to eventually label at least a part of them, it is possible to build automated machine learning models that learn how to classify content and, more in general, solve different tasks simultaneously (multi-task learning [451]). A relevant acceleration in this regard is the ease with which the internet protocols enable users to provide feedback (both explicit and implicit) about their preferences and the degree of the fitness of each item that is provided to them. For example, simply buying or browsing an item may be understood as endorsing that product. Moreover, collecting this data in a web-centric environment is effortless for platforms.

Therefore, information filtering is often based on analysing previous user interactions with recommended items because past preferences are usually good indicators of future choices. The core intuition behind the recommendation task is to find dependencies and patterns between items and users' preferences. These relations can be learned in a data-driven manner, and the learned model can provide new recommendations for target users. To synthesize even more, every model learns a target function $f(x)$ that best maps input (such as purchasing history) variables X to an output variable Y (i.e. next product that will be bought).

Training a neural network involves adjusting the weights and biases of the network to minimize the error between the predicted output and the actual output. This is typically done using an optimization algorithm, such as stochastic gradient descent (presented in Section 2.2.5), which iteratively updates the weights and biases in the direction that reduces the error.

During the training process, the neural network is presented with training examples consisting of input data and the corresponding actual output multiple times (namely *epochs*). The network processes the input data through the various layers, using the weights and biases to make predictions. The error between the predicted and actual output is then calculated using a loss function, such as mean squared error.

The optimization algorithm then backpropagates the error through the network, adjusting the

weights and biases in each layer to reduce the error. This process is repeated for multiple epochs, with the goal being to minimize the overall error across all training examples.

It is essential to carefully select the model's hyperparameters, such as the learning rate, the number of hidden units, the training data, and the loss function, as these can significantly impact the model's performance. The training process can also be regularized using dropout [373] or weight decay techniques to prevent overfitting and improve generalization (which means the performance in predicts new, unseen data, namely the test set.).

Before going deeper into the recommendation task, it is helpful to provide a general understanding of the different tasks that can be solved using machine learning algorithms and their procedure starting from the basic building block of any ML model: the data.

2.2.1 The basic building block of ML: the data

Any ML model requires an input to produce an output, and this input can be represented using vectors, which are mathematical objects that describe euclidean spaces. A vector can be represented as \mathbb{R}^n – *dimensional* space, and instances of the input represented using vectors can contain N – *dimensions*, and each one of them is one attribute (or feature) of that sample. In this sense, vectors are helpful data representations to wrap all the independent variables (inputs) in a single object that can be fitted into the model. ML models can ingest a vector of pixel values from an image. This vector could be represented more formally in this way:

$$\mathbf{X} = [x_1 \quad x_2 \quad \dots \quad x_n] \quad (2.1)$$

where \mathbf{X} is the input vector and x_1, x_2, \dots, x_n are the individual elements of the vector. Usually, these individual values are floating numbers, but the scales of different features can impact learning. To avoid the possibility that the different scales deteriorate the learning performance, other solutions can be adopted [255].

- **Scaling:** allows to restrict the range of data between $[0, 1]$ the formula applied is this:

$$X = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2.2)$$

The advantage is that scaled variables will be in the same range but with different variances. The mean will also not be equal for all the variables differently from the Standardisation.

- **Standardisation:** the input is transformed into using the Z – *score*. The mean of the output will be 0 and the variance 1. The standardized feature is computed using the following formula:

$$X = \frac{x - Mean_x}{Std_x} \quad (2.3)$$

The advantage of this method is that it *centers* the variables around the same value.

The aforementioned methods to normalize data assume that data distribution (between training and testing dataset) and its moments (mean and variance) are fixed in the dataset. The hypothesis behind this is the **stationarity** that implies at least a fixed first and second momentum of the distribution.

This is particularly relevant when dealing with time series. Statistical tests exist that can check for the hypothesis of stationarity and eventually find unit roots such as the *augmented Dickey-Fuller test* [82]. If this is the case, one viable solution could be:

- **Differencing**: time series is transformed into a new one applying the differences between consecutive observation values. *First differences* correspond to one difference. In contrast, *second differences* correspond to two times the difference between consecutive values if even the *first-order-differenced* times series are still affected by non-stationarity.

Tabular data, images, audio, and texts fall under the category of **Euclidean data** that can be represented using vectors, and it is possible to plot them in a linear space. In Euclidean data, the distance between any two points can be calculated using the Euclidean distance formula. This formula calculates the distance between two points in a Euclidean space based on their coordinates. For example, in a two-dimensional space, the Euclidean distance between points (x_1, y_1) and (x_2, y_2) is given by the following formula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2.4)$$

On the other hand, **non-Euclidean** data does not have a fixed distance formula. In a social network, for example, the distance between two people might be based on the number of connections they have in common rather than on their coordinates in a Euclidean space. Non-Euclidean data are often represented using graphs, where nodes represent individual entities and edges define relationships between those entities. The distance between two nodes in a graph is often calculated using a non-Euclidean distance measure, such as the shortest path between the nodes.

But as the geometry, the real world is not only made of Euclidean data. If a line is drawn on a sphere, they will be parallel, but they will eventually cross each other at the pole, and this violates a fundamental axiom of Euclidean geometry. Graphs are widely used data representations for non-euclidean data structures such as social networks or molecules. Algorithms such as Graph Neural Networks (see chapter 6 for examples) seek to adapt existing methods to process non-Euclidean structured data as input directly. The two ingredients for graph networks are:

- **Adjacency matrix** to store the edges. The adjacency matrix is a square matrix used to represent a finite graph. The rows and columns of the matrix correspond to the graph's vertices, and the element in the i_{th} row and j_{th} column is 1 if there is an edge between vertex i and vertex j , and 0 otherwise.
- (To handle attributed graph) **Input feature matrix** $N \times F$, where N is the number of nodes and F is the number of input features for each node. The same applies to weighted edges.

In Figure 2.2, an attributed directed graph is reported, where only nodes have attributes, but in Section 4.3.1, more detailed coverage of graph theory is provided.

2.2.2 From data to Dataset

In ML, it is common practice to divide the available data into training and test datasets. The training dataset is used to fit the model's parameters, while the test dataset is used to evaluate the model's generalisation capabilities. There are several reasons why this is done.

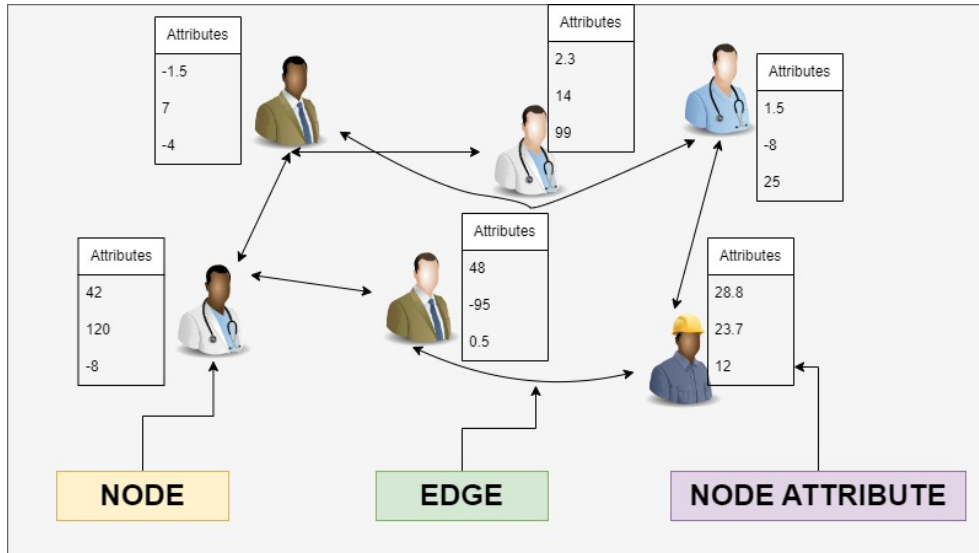


Figure 2.2: In the picture, a representation of a social network as an attributed and directed graph is depicted

1. Using a separate test dataset allows us to get an accurate estimate of the model's generalization error. The generalization error is the difference between the model's performance on the training data and the model's performance on unseen data. Suppose the model is only evaluated on the training data. In that case, the model's error could be a misleadingly low estimate rate because the model has already seen and learned from this data. Evaluating the model on the test dataset makes it possible to get a more accurate estimate of how the model will perform on new, unseen data.
2. Dividing the data into a training and test dataset allows us to tune the model's hyperparameters. Hyperparameters are the parameters of the model that are not learned from the data, such as the learning rate or the number of hidden units in a neural network. The training dataset is used to fit the model's hyperparameters, and then on the test dataset, the model is evaluated to see how well it performs. This allows for finding the optimal set of hyperparameters for the model.

It is essential to carefully consider how to split the data into a training and test dataset. First of all, to evaluate model generalisation capabilities, data overlap between training and test must be avoided because, in that case, the test results will be overly optimistic, as the model was using the same data during training already. The final model performance might be too high. If the data is not representative of the task, the model's performance may be poorly estimated. For example, if the data is highly imbalanced, with a disproportionate number of samples belonging to one class, then the model may be biased towards predicting the majority class. In this case, it is important to stratify the split, ensuring that the proportion of each class is preserved in both the training and test datasets.

In summary, dividing the data into a training and test dataset is a crucial step in machine learning because it allows us to estimate the model's generalization error and tune the model's hyperparameters. When working with time series data, data must be split into a training and a test dataset. However, there are some additional considerations to keep in mind when working with

time series data. One thing to consider is the temporal dependencies within the data. Time series data is usually correlated over time, meaning that the value at a particular time point is likely to be influenced by the values at previous time points. It must be avoided that the test dataset contains any information that could have been used to predict the training dataset.

One way to split time series data is to use a fixed split, as with other data types. For example, it is possible to use **the first 80%** of the data for training and **the remaining 20%** for testing. However, this method has the disadvantage of potentially introducing a "*training gap*" between the end of the training dataset and the beginning of the test dataset. This can be problematic if the data exhibits long-term trends or seasonal patterns, as the model may not accurately predict future values based on the training data alone.

Another method that can be used to split time series data is called rolling cross-validation. In this method, the data is divided into several non-overlapping folds. The model is trained and evaluated multiple times, with each fold being used as the test dataset once. However, unlike traditional cross-validation, the folds are not fixed; instead, they are "rolled" forward through the time series, with the test fold moving one time step forward at each iteration. This allows the model to be trained on various periods, helping to prevent the training gap problem.

In conclusion, when working with time series data, it is vital to split the data into a training and test dataset in a way that respects the temporal dependencies within the data.

2.2.3 The problems

Machine learning models can learn based on different methodologies with different goals. These techniques can solve different tasks, such as classifying content or mapping languages into vectorial spaces to allow automatic language translation. Below the main problem sets are reported. The division is inspired by the structure of input and output:

1. **Supervised learning (SL)**: based on input and labelled output. Given an optimisation algorithm, it optimizes the objective function based on the computed loss that can be intuitively understood as the difference between the real and predicted value.
2. **Unsupervised learning (UL)**: algorithms are used to investigate and group unlabeled datasets. Such models can uncover patterns. The classic example is clustering: given the sample, UL models output a partition into K clusters such that each sample that belongs to a cluster M is more homogeneous and similar to points from its cluster than with samples from different cluster Q [289].
3. **Reinforcement learning (RL)**: This method falls between supervised and unsupervised learning. It does not have labels; instead, these models can learn about rewarding behaviours that are considered good (in the sense that they provide greater rewards) and/or punishing undesired ones (that provide lower rewards). The goal of the RL system is to maximize the reward in the long run, [388]. RL agents can generally understand the environment and learn through trial and error (through an optimisation algorithm). The main components of RL models are (i) the RL agent, (ii) the environment the agent interacts with, (iii) the policy that the agent follows to take action, and (iv) the reward that the agents earn after taking each action.

4. **Self-supervised learning (SSL)**: learns with labelled data during training but these models, during inference, work as unsupervised models. The general method of self-supervised learning is to learn any unobserved or hidden part of the input from any observed or unhidden part of the input. The typical example is an auto-encoder, a network trained to (i) construct a lower dimensional representation of the input, the so-called Encoder and then (ii) reconstruct the input, the Decoder.

The task:

generally speaking, two types of tasks characterise supervised learning:

- **Classification**: each input sample belongs to one or more (in this case, it's named multi-classification). The goal of the model is to predict the correct class(es) for each sample. More practically, the model's output will be a probability distribution over the classes.
- **Regression**: each sample dependent variable is a vector composed of numbers. When the dimension of the output vector is > 1 , that is the case for multi-output regression.

2.2.4 How to evaluate: Loss functions

In machine learning, loss functions are used to measure the difference between the predicted values and the actual values in a dataset. This difference, or error, is used to train the machine learning model so that it can learn to make more accurate predictions. Many different loss functions can be used for regression and classification problems. Some of the most commonly used loss functions are:

Mean Squared Error (MSE)

This is a commonly used loss function for regression problems. The formula for MSE is given:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.5)$$

where y_i is the actual value, \hat{y}_i is the predicted value, and n is the number of samples in the dataset. MSE measures the average squared difference between the predicted and actual values and is often used to train models that predict continuous variables. In a multi-output regression problem, multiple target variables are predicted. In this case, it is possible to extend the mean squared error (MSE) metric by taking the mean squared error for each target variable and then the average of these errors. This can be represented using the following formula:

$$Multi_{MSE} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{K} \sum_{j=1}^K (y_{ij} - \hat{y}_{ij})^2 \right) \quad (2.6)$$

where N is the number of samples, K is the number of target variables, y_{ij} is the true value of the j th target variable for the i th sample, and \hat{y}_{ij} is the predicted value of the j th target variable for the i th sample.

Binary Cross Entropy Loss

Binary cross-entropy loss is used in binary classification tasks where the goal is to predict the probability of an example belonging to the positive or negative class. The formula for binary cross-entropy loss is as follows:

$$BCE = -\frac{1}{N} \sum_{n=1}^N [y_n \times \log(h_\theta(x_n)) + (1 - y_n) \times \log(1 - h_\theta(x_n))] \quad (2.7)$$

Where:

- N is the number of training examples;
- y_n is the target label for the training sample n ;
- x_n is the input sample n ;
- h_θ is the neural network model with weights θ .

The binary cross-entropy loss penalizes the model for predicting probabilities far from the actual labels. A model that accurately predicts the actual labels will have a low binary cross-entropy loss, while a model that performs poorly will have a high binary cross-entropy loss.

Kullback-Leibler-Divergence

The Kullback-Leibler (KL) divergence loss function measures the difference between two probability distributions. It essentially captures the information loss between ground truth distribution and predictions and can be interpreted as a slight modification of the formula for entropy. It is often used in ML to measure how well a model's predictions align with the actual distribution of the data. In mathematical terms, the KL divergence between two distributions, P and Q , is defined as:

$$D_{KL}(P|Q) = \sum_{i=1}^n P(x_i) \log \frac{P(x_i)}{Q(x_i)} \quad (2.8)$$

Here, P represents the actual distribution, and Q represents the predicted distribution. The KL divergence loss is non-negative and is equal to zero if and only if the two distributions are equal. This means that the KL divergence loss can be used as a loss function to train a model to fit the actual distribution of the data.

Overall, the choice of loss function will depend on the specific characteristics of the dataset and the requirements of the machine learning task. Different loss functions may be more or less suitable for different types of problems, and it is essential to carefully select the appropriate loss function for each machine learning task.

2.2.5 How to learn: Optimizer

In the context of neural networks, an optimizer is a method or algorithm used to adjust the parameters of the network to minimize the loss function, epoch after epoch. The goal of training a neural network is to find the set of parameters that results in the lowest possible loss on the training data, and the optimizer is the mechanism by which this is achieved. During the training process,

the optimizer iteratively updates the network's parameters to reduce the loss. An introductory overview of the different optimisation algorithms can also be found in [346].

Gradient descendant algorithm

Gradient descent (GD) is a famous optimization family of algorithms that train deep-learning models. It is an iterative method that starts with random initial values for the model parameters and updates them in a way that reduces the loss (cost) function, which measures the model's error.

This is done by calculating the gradient of the loss function for each network parameter and then using this gradient information to update the parameters to reduce the loss. The simplest form of GD is the **batch gradient descent**,³ which is defined as follows:

$$\theta_t = \theta_{t-1} - \eta \nabla_{\theta_{t-1}} J(\theta_{t-1}) \quad (2.9)$$

Here, $J(\theta_{t-1})$ is the loss (cost) function, θ_{t-1} are the model parameters at time step $t - 1$, and η is the learning rate, which determines the size of the step that the algorithm takes towards the local minimum of the loss (cost) function.

One of the main disadvantages of gradient descent is that it can be slow to converge, especially for complex and large-scale problems. Additionally, it can also be sensitive to the choice of the learning rate. Suppose the learning rate is not set correctly. In that case, the algorithm may not converge to the global minimum of the loss function. It will converge to a suboptimal (local minimum) solution, depending on the batch size and the number of training epochs.

SGD optimizer

Stochastic gradient descent (SGD) was proposed by [336]. It is a variant of gradient descent that calculates the gradient of the loss function for that set of parameters using a small, randomly selected batch of examples from the training data. The parameters are then updated using the calculated gradients according to the following formula:

$$\mathbf{p}_t = \mathbf{p}_{t-1} - \alpha \mathbf{g}_t \quad (2.10)$$

Where \mathbf{p}_t is the vector of network parameters at time step t , \mathbf{g}_t is the gradient (computed on the sampled batch) of the loss function with respect to the parameters at time step t , and α is the learning rate.

SGD has several advantages over batch gradient descent, which calculates the gradient using the entire training set for each batch. One of the main advantages is that it is computationally efficient, as it only requires the gradient to be calculated on a small subset of the training data. This makes it well-suited to large-scale training tasks where the entire training set may not fit in memory. Another advantage of SGD is that it has more chances to escape from suboptimal local minima in the loss function, thanks to the randomness introduced by sampling the training examples. In contrast, batch gradient descent is more likely to get stuck in such suboptimal minima, as it takes a more deterministic approach to parameter updates.

³A **batch** is a sample of training data fed to the models one by one. Batch are usually equally sized. Once one batch is forwarded, the loss and the gradient are computed and the weights updated.

However, one of the main drawbacks of SGD is that it can be sensitive to the choice of the learning rate hyperparameter. If the learning rate is set too high, the training process may diverge, while the training process will be slow if it is too low. This can make it challenging to achieve good performance with SGD, as the learning rate must be carefully tuned.

ADAM Optimizer

ADAM (Adaptive Moment Estimation) [212] is a popular optimization algorithm commonly used in training neural networks. It is a variant of stochastic gradient descent that uses moving averages of the parameters to provide a running estimate of the second raw moments of the gradients; the moving averages (called momentum) are then used to update the parameters in a way that attempts to compensate for the lack of knowledge about the true second raw moments. ADAM has several critical advantages over stochastic gradient descent. One of the main advantages is that it requires less tuning of the learning rate hyperparameter. In contrast, ADAM automatically adapts the learning rate for each parameter, which can help the training process converge more quickly and reliably.

Another advantage of ADAM is that it incorporate momentum, further accelerating the model's convergence. Momentum is a technique that can help the optimizer to avoid getting stuck in suboptimal local minima by adding a fraction of the previous update to the current update. This can help the model make more significant updates early in training and minor later, more fine-tuned updates as training progress. ADAM can help to reduce the need for manual tuning of the learning rate and can incorporate momentum to accelerate the convergence of the model further.

RMSprop

The root mean square propagation (RMSprop), an unpublished adaptive learning rate optimizer proposed by Geoff Hinton, is an optimizer for training deep learning models. It is an update to the classic gradient descent algorithm that uses a moving average of the squared gradients to scale the learning rate. This helps the algorithm converge faster and more accurately. The RMSprop algorithm is defined as follows:

$$g_t = \nabla_{\theta_t} J(\theta_t) \tag{2.11}$$

$$r_t = \rho r_{t-1} + (1 - \rho) g_t^2 \tag{2.12}$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{r_t + \epsilon}} g_t \tag{2.13}$$

Here, $J(\theta_t)$ is the cost function, θ_t are the model parameters at time step t , η is the learning rate, ρ is a hyperparameter that controls the decay rate of the moving average, r_t is the moving average of the squared gradients, and ϵ is a small constant added for numerical stability.

One of the main advantages of using RMSprop is that it can help the algorithm converge faster than other optimization algorithms, such as gradient descent.

One of the main disadvantages of using RMSprop is that it can be sensitive to the choice of the hyperparameters, such as the learning rate and decay rate. RMSprop can also be computationally

expensive, as it requires keeping track of the moving average of the squared gradients for each parameter.

2.3 Overview on Recommender Systems

On September 21th, 2009, Netflix, a semi-unknown video streaming company, made a bank transfer of exactly US \$1,000,000 to a group of data scientists called the *BellKor's Pragmatic Chaos team*. This team over-performed Netflix's proprietary algorithm to predict users' ratings by 10% and won the Netflix Prize. Thirteen years later, Netflix, leader streaming company around the world, in its official site ⁴ states that: "Recommendation algorithms are at the core of the Netflix product."

It's quite unexpected considering efforts to produce movies and TV shows (€300M invested only in Italy during 2017 – 2020 ⁵) and nowadays, not only one model is used, but rather a set of techniques are employed based on the different recommendation tasks [374].

What is quite not surprising is that today the recommender system "influences choice for about 80% of hours streamed at Netflix [...], and we think the combined effect of personalization and recommendations save us more than \$1B per year." [153]

Recommender systems are an instance of information filtering systems. They are critical in social media because they help to surface relevant content for users in a sea of information. Social media platforms generate vast amounts of data about user behaviour, preferences, and interactions, and recommender systems can use this data to identify patterns and make informed recommendations about what content a user might be interested in. RS can be powered by machine learning models but also with deterministic techniques.

In addition to improving the user experience, recommender systems can also help to increase engagement on social media platforms. By showing users the content they are more likely to find interesting or relevant, recommender systems can keep users on the platform longer and encourage them to interact with more content. This can help to drive user growth and increase the platform's advertising revenue.

Several approaches can be used to solve recommendation tasks, which predict what items a user might be interested in based on their past behaviour or other information about them. One approach is collaborative filtering, which involves building a model based on the past behaviour of a group of users and using that model to make recommendations to individual users. This approach identifies users with similar interests or preferences and uses those similarities to make recommendations. Another approach is focused on content, which involves calculating the similarity between different items or users based on certain features or characteristics and using that similarity to make recommendations. Another approach that has gained popularity in recent years is using neural networks to solve recommendation tasks. In the context of recommendation tasks, neural networks can be trained to learn the relationships between different items and users and to use that knowledge to make recommendations. One advantage of using neural networks for recommendation tasks is that they can handle vast amounts of data and learn complex relationships between different items and users in applications where collaborative filtering can be computationally expensive.

⁴Source: <https://research.netflix.com/research-area/recommendations>, on July 5th, 2022

⁵Source: <https://www.ilsole24ore.com/art/netflix-quote-imposte-d-investimenti-italia-mettono-rischio-sistema> on April 7, 2023

2.3.1 Tasks and Evaluation Metrics for Recommender Systems

Information filtering tasks can be framed in multiple ways to find and serve relevant information to the user. The task can be defined based on the objective:

- **Ranking:** given a set of candidate items for a given query, the algorithms must serve them to the user, starting from the most relevant
- **Preference prediction:** based on the user model, the system should return the items that best match the current user model.
- **Link prediction:** applied in the case of users' recommendation this class of RSs must predict which is the probability that two nodes in the graph (it could be a bipartite graph with users and items or a social network) will create a new edge between them.

The basic assumption in a recommender system is that a system that provides more accurate predictions will be preferred by the user [162]. Starting from this, the task design and the particular aspect of interest for a specific metric can be used to evaluate or optimize the model. The metric of interest depends on the specific domain of the recommended (i.e. a paper recommender system will prioritize items similar to the ones previously browsed, while a news feed recommender system should be optimized to favour a broader diversity and coverage of the set of items). Below is a list of the different aspects the metrics can address in the RS field.

- **Statistical Accuracy:** accuracy is the most used (and debated) metric applied to RS. It focuses on the number of times the RS can correctly make recommendations. It measures the percentage of recommended items that are consumed or interacted with by the user. This can be expressed mathematically as:

$$\text{Accuracy} = \frac{\text{Number of relevant recommendations}}{\text{Total number of recommendations}} \quad (2.14)$$

Here the term *relevant* means that recommendations are coherent with the objective of the system, in the majority of the cases, an engagement metric. These metrics can take many forms, including precision, recall, and mean average precision. In the case of ranking, they can be adapted, and for example, the **Precision@K** is the fraction of relevant items in the top K recommended results.

- **Novelty:** this is a crucial aspect of RS because if users themselves would have easily found the RS-served content, the RS decreases its value and utility for the user. Following [411], novelty generally refers to how different recommendations are with respect to *what has been previously seen*. Novelty can be defined by different sides, such as the distance between the item and the context of the user's experience. Novelty is also related to **Serendipity** that measures the extent to which a user is positively surprised by an item in the recommendations. [141] proposes a serendipity measure that also weights the utility for the user given that although it may be surprising for a user to receive a serendipitous recommendation, this does not mean that it is proper to recommend a completely unrelated item to them.

- **Diversity:** the heterogeneity of recommendations gained attention first because it is a helpful tool for answering ambiguous queries in information retrieval [418]. A typical diversity metric is the entropy of recommended items [226]. Diversity metrics do not evaluate the ranking of items within recommendations. They only consider items' novelty and diversity qualities. Additionally, these metrics do not consider the relevance of the items, only focusing on their novelty and diversity. A concept close to diversity is **Coverage** which is a metric used to evaluate the comprehensiveness of a recommender system's recommendations. It reflects the percentage of items in the system's inventory recommended to users at least once (catalogue prediction [141]). High coverage indicates that the system can provide recommendations for various items. In contrast, low coverage may indicate that the system can only make recommendations for a small subset of the available items. Coverage is similar to diversity in that both metrics focus on the breadth of the recommendations provided by the system. However, while diversity measures the variety of items recommended to a single user, coverage measures the percentage of the overall item inventory recommended to any user. Coverage can be seen as a global measure of diversity, while diversity is a more localized measure.
- **Fairness:** is a concept related to equality. In the context of RS, it can be intended as the equal exposition of items to the users or the equal treatment of individuals that belong to minorities by the RS. Fairness can be crucial in the recommender systems because a biased RS toward minorities can profoundly impact the future of young workers [329]. Fairness can be measured within or between groups of individuals or items [369]. These metrics can include measures of demographic parity, equal opportunity, and counterfactual fairness, which assess the proportion of recommendations received by different demographic groups, the proportion of relevant recommendations received by different groups, and the hypothetical impact of changing the protected attributes of a user on the recommendations received, respectively. A detailed review of fairness is provided by [229] that highlights how *hybrid fairness* can balance different forms of fairness. The possibility to intervene to fix are multiple and can be positioned in the *pre-*, *in-*, and *post-* processing methods [229] to compensate algorithmic or data bias.

Evaluating the performance of an RS is a complex task due to its composite nature and different aspects researchers or industries could be interested in. The classic approach can be summarized in the optimization for *engagement*. This concept can be intended as a proxy for “the probability of desired or targeted user reactions”. This concept highlights how the company's needs are prioritized because this may or may not be equal to maximising user value even if engagement provides some correlation for “value”, but these two concepts don't need to overlap [265]. Moreover, [179] establish an *engagement-diversity trade-off*.

A broader approach to this task is to consider the different "families" of metrics relevant to different aspects of the recommender system's operation. These families include accuracy, diversity, and fairness metrics, among others. It is essential to consider and value the exploration of recommenders improving metrics such as diversity and fairness when designing and evaluating recommender systems because they can impact the user's experience and overall effectiveness, especially in the long term [79].

2.3.2 Collaborative filtering

The collaborative filtering approach is a widely-used method for making recommendations in recommendation systems. This approach uses individuals' ratings or preferences to predict others' ratings or preferences, leveraging latent representation and preferences' redundancy. In other words, collaborative filtering uses a sort of *wisdom of the crowd* to make recommendations by identifying patterns and trends in the ratings or preferences of many individuals.

Mathematically, the collaborative filtering approach can be implemented using matrix factorization. This involves representing the ratings or preferences of individuals as a matrix, where each row corresponds to a specific individual and each column corresponds to a specific item (such as a book, movie, or product).

One way to implement matrix factorization for collaborative filtering is using the Singular Value Decomposition (SVD) algorithm. The SVD algorithm decomposes the rating matrix into three matrices: a user-factor matrix, an item-factor matrix, and a diagonal matrix of singular values. These matrices can be multiplied together to approximately reconstruct the original rating matrix, and the resulting product can be used to make recommendations for individuals.

The formula for the SVD algorithm applied to collaborative filtering is given:

$$A \approx U \cdot \Sigma \cdot V^T \quad (2.15)$$

where A is the ratings matrix, U is the user-factor matrix, Σ is the diagonal matrix of singular values, and V^T is the transpose of the item-factor matrix. The collaborative filtering approach using matrix factorization is practical for making recommendations in a wide range of applications [354].

There are several potential drawbacks to the collaborative filtering approach. One of the main challenges of this approach is the so-called *cold-start* problem, which occurs when there is insufficient data available to make accurate recommendations for a given individual. This can happen when an individual is new to the system and has not yet provided any ratings or preferences or has only provided a small number of ratings or preferences. In these cases, the collaborative filtering approach may not be able to make accurate recommendations and may instead provide generic or uninformative recommendations.

There are several ways to address the so-called "*cold start*" problem in recommendation systems. The most straightforward approach is to use users' demographics, assuming a set of users with similar features have more or less similar preferences. One approach is to use a hybrid recommendation system, which combines the collaborative filtering approach with other methods for making recommendations [362]. For example, a hybrid recommendation system could use collaborative filtering to make recommendations based on the ratings or preferences of similar individuals and estimate ratings using content-based methods such as a k-nearest neighbour. It could also use content-based filtering to make recommendations based on the characteristics of the items themselves [392]. By combining these two approaches, a hybrid recommendation system can provide more accurate and relevant recommendations even when insufficient data are available to make recommendations using collaborative filtering alone. Another approach to addressing the cold start problem is to use transfer learning [137]. This involves training a recommendation model on a large dataset with a high degree of variability and then fine-tuning the model on a smaller dataset with more specific

characteristics.

Overall, there are many ways to address the cold start problem in recommendation systems. The best approach will depend on the dataset's specific characteristics and the recommendation task's requirements. Another potential drawback of the collaborative filtering approach is that it can be susceptible to bias such as the popularity bias [46, 56]. Popularity bias is related to the underlying distribution of ratings among items. Given the presence of a *long tail* in the distribution, which means that many items have few ratings, collaborative filtering RS typically emphasizes popular items (those with more ratings). Even if popular items are often good recommendations, their popularity can reduce the effectiveness of the systems, which can also reduce the diversity and novelty of recommendations. Delivering only popular items may not encourage the discovery of new items and may not consider the preferences of users with niche interests. Additionally, this approach may be unfair to producers of newer or less popular items, as fewer users are rating them. Other issues can emerge. For example, suppose the ratings or preferences of individuals are not representative of the overall population. In that case, the recommendations produced by the collaborative filtering approach may be biased and may not accurately reflect the preferences of the entire population. Finally, collaborative filtering approaches suffer the relatively high computational cost of, i.e. decomposing user-item matrix and the fact that the model must compute the entire matrix at each update of the model. This implies low scalability for the model.

2.3.3 Content-Based

The Content-based approach is a method for making recommendations in recommendation systems. This approach is based on using the characteristics or features of the items and users themselves to make recommendations [288]. The recommendation process involves comparing the characteristics of a user's profile with the characteristics of a content object, resulting in an estimated rating that reflects the user's level of interest in that object. For example, if two items have similar characteristics (such as genre, style, or author), the item similarity approach would recommend one item to an individual who has expressed an interest in the other item. Following [240], the main components of a content-based recommender are:

- Content Analyzer: return a more suitable representation of content, often embedding them into a latent space.
- Profile learner: collect data about the users and return a user model that embeds its preferences.
- Recommender: match the user profile and item representation to provide recommendations.

One of the main drawbacks of the item similarity approach is that it relies on the availability and reliability of the characteristics or features of the items, highlighting the relevance of the content tagging or embedding system. If the characteristics or features of the items are not accurately represented, or if the characteristics or features are not sufficiently descriptive, the item similarity approach may not be able to make accurate recommendations. Concerning the feedback loop in this context, determining what characteristics of the item the user dislikes or likes is not always obvious.

Another potential drawback of the item similarity approach is that it is not designed to provide serendipitous recommendations and promote users' diversity exposure. Conversely, a content-based recommender is also prone to overspecialization: they will recommend items similar to those already consumed, with a tendency to create a *filter bubble* [308] or provide low-quality recommended items.

2.3.4 Hybrid recommenders

A hybrid recommender system combines the strengths of multiple recommendation approaches, such as collaborative filtering and content-based filtering, to provide more accurate and diverse recommendations and mutually solve each other's drawbacks [397], integrating, weighting [45] or combining predictions from different and separate models. Combining the complementary information provided by different recommendation techniques, hybrid recommender systems can often improve the system's overall performance to address several issues that arise with other filtering approaches [65], such as the *cold start problem*, *overspecialization problem*, and *sparsity problem*. Hybrid recommenders can also be based on the *hybridization of representations*, such as combining graph embeddings and contextual word representations [317] to feed a deep architecture that provides recommendations and results confirm the validity of the intuition behind the proposed framework.

2.3.5 Neural Network Recommenders

Neural network-based recommender systems use machine learning to learn the relationships between items and users from historical interaction data. One popular approach to neural network-based recommendation is the use of autoencoders, which are models that can learn the underlying structure of the data through dimensionality reduction. Autoencoders have been used in recommendation tasks by learning the latent features of the items and users and using those features to make recommendations based on the similarity between items or users [382]. Also, [305] proposed an autoencoder-based architecture to learn social representations for a recommender system to address the data sparsity and imbalance problems for social networks.

Another approach is using attention mechanisms, which allow the model to dynamically weigh the importance of different items or users in the recommendation process. This can be particularly useful in cases where the model designer needs to take into account the temporal sequence of the interaction with the recommender (instead of considering static the interaction) as in [441], where the long-term and short-term user-item embeddings are fused in a separate stage of the model.

Graph-based neural networks (GNN) have also been utilized in the context of recommender systems, and [436] provides an extensive review of the field. The main advantage is that they can effectively and naturally capture the complex relationships between users and items in a recommendation task and encode the crucial collaborative signal (i.e., the topological structure of the underlying graph) and also include temporal dynamics [124, 282]. These models operate on graph structures, where the nodes represent the users and items and the edges define the relationships between them. One advantage is that GNNs can manage heterogeneity, creating multiple types of nodes or edges. Ideally, a graph layer also allows to naturally include side information as a graph structure, such as a social relationship and knowledge graph that can enrich the information given to the model. The different approaches proposed to manage the message passing stage (when in-

formation from each node’s neighbourhood is aggregated) allow to create of multiple graph layers such as Graph Convolution [214], Attention (GAT) [414], Gated and recurrent (GRNN) [347].

In conclusion, machine learning and recommender systems are closely intertwined and play a significant role in shaping social media content. Recommender systems powered by machine learning algorithms are capable to analyze vast amounts of user data, including past interactions, preferences, and search histories, to personalize content recommendations. These recommendations are designed to improve user engagement and satisfaction by showing them content that is relevant to their interests and needs. However, the use of machine learning algorithms in recommender systems can also create, under certain conditions, echo chambers and filter bubbles, reinforcing users’ existing biases and limiting exposure to alternative perspectives and information. Understanding the link between machine learning and recommender systems is crucial to developing safe social media and maximising the benefits associated with these technologies. A more constructive approach than questioning whether social media is causing polarization would be to explore the potential of social media interventions to alleviate it, also using machine learning.

2.4 Conclusion

In conclusion, this chapter provided an overview of the basics of machine learning, including supervised and unsupervised models, optimization algorithms, loss functions, and the different types of tasks that can be framed in multiple problem settings. The chapter also covered the fundamental knowledge about recommender systems, including content-based, collaborative filtering, and hybrid recommenders, and a comprehensive overview of recommender evaluation methods beyond accuracy, which account for fairness, diversity and coverage. With this comprehensive foundation in place, the reader is now equipped with the tools necessary to understand social media and recommender biases and potential threats, which are addressed in the next chapter.

Chapter 3

Navigating Social Media Risks and Recommender System Biases: A Case for Human-Centered AI

This chapter examines the potential biases and threats in social media, in particular in recommender systems. In the second half of the chapter, the topic of AI and human value alignment is considered, focusing on the ethical implications of artificial intelligence and the potential consequences of misaligned values. Finally, the possibility of building a collective well-being-oriented social media platform that addresses these issues and promotes the greater good is reflected upon, and the notions of the **educationally managed social media community and digital collective well-being** are introduced.

3.1 Social Media Threats classification

It can be argued that social media are, like society, a complex system because it involves a considerable number of elements arranged in structure(s) with multiple and multi-modal interactions that exist on many scales and are not reducible to only one level of explanation [161].

Social Media data are gathered as sequences of events that are observations of various complex dynamic processes, characterized by the fact that are mostly noisy text or video, in an unstructured manner, with a short length, where stylistic variations, acronyms and slang are common, and where context plays a crucial role in determining the meaning and relevance of the content. Therefore, extracting valuable insights and meaningful patterns from social media data requires sophisticated natural language processing, machine learning, and data mining techniques [33, 41].

Understanding the composition and nature of what can be named "**social media threats**" is needed. The fact that multiple stakeholders (such as users, creators, institutions, and news companies) are involved in these multi-sided platforms anticipates that multiple levels coexist where these threats can originate.

In the following sections, different levels and sources of social media threats are highlighted, and this will set the ground for the chapters presented later in this section.

3.1.1 Individuals: cognitive limits and emotional state

Questioning the legitimacy of content shared on social media is relatively easy if legal criteria or journalistic approaches are given, but automatic content moderation still has several limitations [236]. Still, it is more challenging to state which individual traits lead to sharing or believing in questionable or harmful content and to what extent users are aware of internal and external factors that influence their behaviour.

It could be argued that social media users can be considered a threat to themselves because of their cognitive limits and biases. Still, also their emotional state can negatively impact their behaviour. Together, these factors can make users vulnerable to manipulation, exploitation, and self-harm, and the crucial question is how to reduce these risks without suppressing useful content [183]

Different limits characterize the user [427]. Still, above all, there is the fact that the information received exceeds people's limits in terms of cognitive capabilities to process all the information social media produce. This has led to an increased focus on the **economy of attention**, as theorized by Simon [367], in driving information dynamics.

Users can experiment with different types of overload in terms of system features, information, or social overload [39, 134] that can affect the expected reaction to information pieces.

Moreover, users characterized by different cognitive styles can be more susceptible to misleading content [16, 313]. Personality traits can be associated with different social media usage [180], which also relates to different problematic smartphone usage dynamics with different personalities. The emotional state also plays an essential role because it can be transferred through the network between individuals [219] without a clear and intentional purpose. Emotions and moral messages in information pieces (such as texts) can also shape the content's diffusion on social media, increasing the retweet rate by around 20% [59].

Individual political slants can affect content evaluation, and adults have a clear tendency to believe in stories that favour the preferred political party, especially in the case of segregated networks [16] where a strong correlation between connected users is found.

People also tend to avoid cognitive discordance when acquiring new pieces of information and only assimilate confirming claims [104, 316, 328]. Moreover, another tendency of human behaviour framed by sociologists is the so-called *homophily* which indicates the willingness to create bonds with like-minded people. These two factors together have the potential to create homogeneous clusters of opinion [423]. Empirically, Garret *et al.*, [140], highlights that people tend to seek reinforcing opinion news items and spend more time reading them.

Teenagers are also prone to problematic smartphone usage can degenerate into digital addiction, characterized by compulsive usage with clear harm to mental and physical state [18]. Addiction by non-pharmacological factors [295], such as social media, can be a crucial factor in problematic smartphone usage [224].

Moreover, the lack of non-verbal communication and limited social presence [163, 259, 341] often exasperates carelessness and misbehaviours, as the users perceive themselves as anonymous [109, 319], do not feel judged or exposed [430] and deindividualize themselves and other users [244].

As will be detailed later on (see Section 5.1.1), the relevance of education to digital platforms, the so-called *Digital Literacy* introduced by [148] time before social media advent is becoming a valuable tool to counteract social media threats.

Users' offline behaviour can also be deteriorated by online behaviour and their perception of online life. Phenomena such as the **Fear-of-Missing-Out (FoMo)** refers to anxiety around missing out on rewarding experiences that others are having, keeping people looking for new content and interactions (and as a consequence, the rewarding) that can alleviate social exclusion that the user is feeling. FoMo can also have a role as a predictor of problematic smartphone usage [49, 131].

3.1.2 Harmful content characterisation

Concerning content, the first challenge that digital platforms have to face is the rise of multi-modal content that combines different modalities inside one *post*¹: texts, images, short videos and external links can compose a single message, and there is evidence that the union of two different modalities (for example, an image placed next to a text) can affect the perceived veracity of the message [284], and more in general, nonprobative images (or words) can inflate the perceived accuracy.

Harmful content can be characterized apart from the modality based on the different targets and scope of the content. The most common type of harmful content is the so-called *fake-news*. This broad term indicates a piece of information (usually a text composed of multiple claims) that is verifiably false and could mislead readers. The willingness to diffuse such content is also a relevant characteristic, and these main features help to map into an initial and broad categorisation of these different types of *fake news*:

- **Misinformation**: incorrect or misleading information.
- **Disinformation**: A form of propaganda involving disseminating false information with the deliberate intent to deceive or mislead.
- **Malinformation**: the deliberate publication of private (and genuine) information for personal or corporate rather than public interest, such as revenge porn.

Apart from content that can be characterised as *news* related to political agenda, other examples of harmful content with a specific target can include **beauty stereotypes** [416]. This is a typical feature of social media content due to several reasons. Firstly, social media platforms are designed to encourage and amplify the sharing of visual content, which includes images and videos that promote narrow beauty ideals. Users are constantly exposed to images of seemingly perfect bodies and faces, which can lead to a negative impact on their self-esteem and body image. Secondly, social media algorithms tend to prioritize and promote content that is likely to generate engagement and interaction, such as images that conform to idealized beauty standards, because they can fit the preferences of a broader audience. This can create a feedback loop where users are more likely to see and engage with content that reinforces body stereotypes, leading to further amplification of such content. Thirdly, social media platforms are also used as a means of self-presentation and self-promotion, with many users using their profiles to showcase their bodies and appearance. This

¹In social media, a post refers to a piece of content that a user shares on their profile or page. This content can take many forms, including text, images, videos, and links. Posts can be seen and interacted with by the user's followers or friends, depending on the platform and the user's privacy settings. Users can also interact with posts by liking, commenting, or sharing them. Posts are the main way that users communicate and share information on social media, and they are often used to share personal updates, thoughts, and experiences, as well as news, articles, and other content.

can create social pressure to conform to narrow beauty standards, leading to further promotion and perpetuation of body stereotypes.

Also, **hate speech**, which can be better characterized and studied in the different hate narratives that can take place with different declination: race, ethnicity, and religious-based discrimination. This is particularly relevant because it is the first threat to be legally coded under the umbrella of hate crimes that can take place on social media, which should be addressed as one of the faces of the systematic racism rooted in almost all countries around the world.

As mentioned before, multi-modal content and images dominate the arena in terms of interaction. It is not only a commonplace that a *"A picture is worth a thousand words"*. Images can convey multiple concepts more succinctly. Beauty stereotypes can be conveyed through social media [416] because the images posted are used by teenagers to estimate their self-worth and correspond to the projection of idealized beauty standards onto the posted content [85]. Still, images also need more context to be understood. Given the increased cognitive load needed to correctly couple with the fact that recent models can create realistic images, they focus particular attention on images [257].

The algorithms that fall under the term *generative models* are known for their capability to synthesize realistic-looking data that can threaten users if not correctly interpreted. More formally, these models learn to estimate the underlying distribution of parameters that govern actual data generation and maximize the likelihood that synthetic data resemble real ones. In particular, the class of Generative adversarial networks [154] can generate newspaper articles written by AI or images that look like real photographs or paints [330]. Generative models can also be applied to text and, based on this dataset composed of 3.5 years of 4Cahn posts[307], Yannic Kilcher (a famous YouTuber) trained a language model ² reaching performances similar to small GPT-models regarding the benchmark **TruthfulQA** [232]. OpenAI a company that deploys AI models released on the 30th of November 2022 a Large Language Models (LLMs) called ChatGpt. This model is a type of artificial intelligence that has been gaining attention in recent years. These models (i.e. also Meta released a LLM) are designed to process and generate natural language text using complex algorithms and deep neural networks. LLMs can learn from vast amounts of data, making them highly efficient and effective at tasks such as language translation, text summarization, and language generation. The main advantage of LLMs is their ability to generate high-quality text with very little human intervention. They can analyze and understand large amounts of data, and then use this information to generate coherent and meaningful text. This makes them highly useful for tasks such as automated content generation, chatbot development, and even language learning. However, LLMs are not without their limitations. One of the main challenges is their requirement for vast amounts of data to be trained effectively. This can make them computationally expensive and time-consuming to develop. Additionally, LLMs have been criticized for their potential to generate biased or offensive content, especially when they are trained on data that contains such biases. Despite these challenges, LLMs have proven to be a highly valuable tool in natural language processing, and their development is likely to continue to advance in the coming years. Future research will likely focus on developing more efficient and effective LLMs, as well as addressing the potential ethical concerns surrounding their use.

²Video Source here <https://www.youtube.com/watch?v=efPrtcLdcdM>, code source here <https://github.com/yk/gpt-4chan-public>

3.1.3 A solution to harmful content: Threat Detectors and Content Analyzers

The great variety of social media threats (as described in Section 3.1) results in challenging issues, and researchers are studying how to identify them automatically. Still, this path opens perils for the freedom of speech because also the moderation policy itself can affect linguistic and self-censorship [144]. One way of bringing together the researchers' community to work on solving social media threats is workshops on these topics, e.g [222, 281]. As introduced in the beginning, another way is challenges (or shared tasks). Examples include hate speech detection at HaSpeeDe in Evalita 2020 [352] or toxic span detection at Semeval-2021 [310].

Solutions proposed to counteract threats on social media are usually defined as classification tasks commonly solved using deep learning. Depending on the type of threat, the input can include textual, visual or network signals. Methods and models developed as part of the COURAGE project are presented and used to detect threats in the proposed framework. This includes **(1) classifying textual content**, **(2) analyzing visual content** and **(3) revealing network structures like echo chambers**. The general architecture is flexible so that new classifiers can easily be added or replaced in a plug-and-play fashion.

3.1.4 Text-Based Detectors

With a vast amount of social media threats taking a textual form, text-based detectors are presented and categorized by different threats.

Hate Speech and Toxic Content

An approach to profiling hate-speech spreaders on Twitter was submitted to CLEF2021 and features runs for multiple languages [10]. For English, a pretrained BERT model was fine-tuned, while for Spanish, a language-agnostic BERT-based sentence embedding model without fine-tuning was used.

Transformer models are widely adopted in solving text classification tasks, and [178] use them to generate text representations for their submission at the Evalita 2020 shared task on hate speech detection.

Transformer models for hate speech detection were also used for identifying irony in social media [406]. Ensembles of transformer models and the automatic augmentation of training data were proposed. Using the standard SemEval 2018 Task 3 benchmark collection, they demonstrate that such models are well suited in ensemble classifiers for the task at hand.

However, also other methods are introduced, for example, an approach based on graph machine learning by [431]. The participation in the Hate Speech and Offensive Content Identification (HASOC) [270] campaign aimed at examining the suitability of Graph Convolutional Neural Networks (GCN) due to their capability to integrate flexible contextual priors as a computationally effective solution compared to more computationally expensive and relatively data-hungry methods, such as fine-tuning of transformer models. Specifically, the combination of two text-to-graph strategies based on different language modelling objectives was explored and compared to fine-tuned BERT.

Another graph-based method in the context of hate speech detection, more specifically sexism detection, was introduced in [432]. This method builds on Graph Convolutional Neural Networks (GCN), exploring different edge creation strategies and combining graph embeddings from different

GCN through ensemble methods. In addition, different GCN models and text-to-graph strategies are explored.

Despite the success achieved by these efforts, the robustness of these systems is still limited. They often cannot generalize to new datasets and resist attacks (for example, word injection) [159, 181]. Some recent models can generalise the task while maintaining similar results in different platforms and languages under certain conditions [432]. In general, this is important as small changes impact the system performance, making it challenging to apply these approaches in the dynamic contexts of social media.

Fake News and Misinformation detectors

To detect fake news, an approach was proposed in the context of the Conference and Labs of the Evaluation Forum (CLEF2021). It applies automatic text summarization to compress original input documents before classifying them with a transformer model [169]. Promising performance was reported on the utilized dataset, while the system has also established a new state-of-the-art benchmark performance on the commonly used FakeNewsNet dataset [170].

Other recent methods apply ensembles of different models for fake news detection [11, 402]. While the focus in [402] lies on transformer-based models to predict the labels, [11] also combines these with a support vector machine. In both papers, the dataset from the GermEval2021 shared task was used.

In general, fake news detection datasets have frequently been proposed as part of shared tasks, and they are used as, for example, in [403] or [238]. While [403] apply automatic text summarization, similarly as in [169, 170], and combine this information with automatic machine translation, [238] introduce an approach that is based on text graphs and graph attention convolution. Although submissions were very competitive, the contributions by [403] demonstrate that this approach is highly competitive as they resulted in winning the German cross-lingual fake news detection challenge at CLEF 2022 "CheckThat!" [403].

User Beliefs and Opinions detectors

Models to extract user-related properties are popular, such as beliefs and opinions or sentiments and emotions. Inferring and interpreting human emotions [318] includes distinguishing between sentiment analysis, the polarity of content (e.g. [164, 165, 235]) and emotion recognition (e.g. [8, 40]). In comparison, opinion extraction aims at discovering users' interests and their corresponding opinions [422]. Similarly, the positive aspects of social media interaction, crucial for estimating the "*collective social well-being*", could be extracted. Still, they have attracted less attention, but see [78, 421].

The first step needed is to educate users about those types of content and how to inspect content and get more information if the provided context is insufficient to clarify the main aspects.

3.1.5 Auditing Algorithms

Multiple scholars have recently addressed algorithms' drawbacks and auditing processes have been proposed, given the relevance mentioned above of AI concerning companies' needs and goals and users' satisfaction or utility in the long term. As stated by [4], the AI agenda should prioritize

the development of "*Explainable, Accountable and Intelligible Systems*". Algorithmically curated feeds are common but resistance to algorithmic change [106] largely revolves around expectation violation but this highlights the fact that introducing information filtering must be placed side by side with strong monitoring and auditing systems to ensure that models are aligned with the given objective without harming users or creating negative side-effects. Scholars have found that users' awareness of algorithms was low and their level of understanding varied widely [122]. As highlighted by Xie *et al.*, [439], platforms should further improve users' awareness and knowledge formation of algorithms by increasing the visibility of information and functions of algorithms. Along with improving usability and helping users to become more skilled, it is crucial to understand which are the potential algorithms' drawbacks and how to investigate them.

For example filter bubbles and echo chambers are social media-specific threats because they are enabled and amplified by the unique features of social media platforms. Social media algorithms are designed to prioritize and promote content that is likely to generate engagement and interaction, which can lead to the creation of echo chambers and filter bubbles. In an echo chamber, individuals are exposed only to content that reinforces their existing biases and beliefs, leading to the formation of narrow and isolated communities of like-minded individuals. In a filter bubble, social media algorithms create personalized feeds that show users only content that matches their previous interactions, preferences, and search histories. This can create a self-reinforcing cycle where users are exposed only to content that confirms their existing beliefs and biases while being shielded from alternative perspectives and information. Social media platforms are uniquely positioned to create and amplify these threats due to their ability to collect and analyze vast amounts of user data, and their use of AI algorithms to personalize content delivery. As such, addressing the threats of echo chambers and filter bubbles requires a proactive regulatory framework and requires a deep understanding of the real dynamics of these threats and the factors that affect them, starting from the algorithms.

A taxonomy of the auditing process concerning the distortion taken into account, auditing methodologies, and organizations audited has been summarised by [34]. Different strategies can be adopted to test an algorithm as pointed out by [351]:

- **Code access:** training data, source code and evaluation metrics are directly accessed by researchers.
- **Puppet:** data are collected with computer programs that simulate users' behaviours. Tomlein *et al.*[398] simulate users with a preference model for a given topic, which delves into a misinformation filter bubble and then tries to burst the bubble. Results show that bursting a filter bubble is possible. Zhang *et al.*, [450] evaluate a conversational item recommender with simulated users, and findings show that preference model and task-specific interaction models can achieve a high correlation between automatic and human evaluations. Yao *et al.*, [444] build a selection and feedback model to simulate interaction with RS and study how popularity bias manifests in repeated interactions
- **Scraping:** official API³ or un-official web scraping techniques could be used to obtain data.

³an API (Application Program Interface) is a web interface used to ease the communication between different software to exchange data.

- **Crowd-sourced:** real people are involved. [119] proposed a framework to evaluate RS in the context of technology-enhanced learning.

It is necessary to clarify that AI often reflects biases already present in society [38]. People have always loved blockbusters, even before the advent of streaming platforms, and the training data that can be fed to an AI to learn which movies fit customers' preferences will be affected by the fact that a lot will watch a few movies (blockbusters). In contrast, the majority of the films will have few or no views at all. This poses a first issue concerning the diversity recommender systems should offer users. Golden and Danks, [152], address the ethical obligations to provide novelty content and highlight the overlapped and blurred nature of the platform and users' interests. On the one hand, platforms have no moral obligation to society to provide diversified content. Still, at the same time, if novelty affects users' engagement (that means profits) positively more than the cost of this *novelty bias*, there is room to provide more diversified content.

Other studies are pointing out that even if recommender systems could reduce the diversity of content [75], they can at least preserve the diversity of content provided in news recommendation [272].

Obviously, more diversified recommendation strategies should take into account the scope of the platform but also the individual tolerance of novel, unseen and maybe unexpected content to balance the potential serendipity of the system. To preserve at the same time, the "long tail" and personalisation, Abdollahpouri, [3], proposes to add a personalisation factor which values the interest in items that belong to the long tail for each user. This two-factor function increases the coverage of the distribution of both *short* and *long* tail items.

Accuracy, novelty, and diversity are the main aspects of the evaluation of recommended items list in broader terms; in [113], a survey to investigate user perception with respect to the aforementioned aspects reveals that users' satisfaction is influenced positively by diversity rather than novelty which is anyway correlated and this indicates that a little share of unfamiliar (novel) items can improve satisfaction. Still, there is a threshold where users indicate a decreasing utility when too many unfamiliar items are provided.

The relevance of recommender and filtering system in news-feed create a vigorous debate concerning the potential of AI to create a *bubble* that only allows items that fit closely to users' preference to enter of feed. In 2011, Eli Pariser coined the term "Filter Bubble" [308] to describe the fact that recommendation algorithms create for each one of us a unique universe of information that is biased towards our interests and give a unique perspective but also limits the access to the full spectrum of information. Moreover, personalisation puts in danger democracies because it hinders the possibility

to see things from one another's point of view, but instead, we're more and more enclosed in our bubbles. Democracy requires a reliance on shared facts; instead, we're being offered parallel but separate universes [308]

Empirically [286] found that diversity increased for users who followed provided recommendations in a movies' recommendation setting. While [23] highlights that algorithmically-driven listening through recommendations is linked with diminished diversity of streamed tracks. An explanation of these results that seem opposite of each other could be found in the fact that there is

probably a natural tendency to form habits that can be observed in the narrowing consumption of items (movies or songs).

Users could also be influenced by what other users think. In the case of a rating system, users can be manipulated because they tend to rate toward the expected value shown to them by the system [96]. The user-specific factors that affect the recommender system in the most recent and disruptive social media, TikTok, owned by the Chinese company ByteDance have been studied through simulation in [54] with a puppet methodology. Results highlight that conscious factors, such as the following, and unconscious, like video view rate, affect what is recommended to users. It is not transparent with this methodology which is the impact of the initial state of the users (the synthetic distribution of its features such as hashtags and creators interested in) concerning the diversity of provided content.

As mentioned before, recommender systems are not only in charge of filtering information and content, but also they suggest new people to follow and could learn to reflect harmful dynamics. It is well known that minorities are sometimes exposed to the so-called *glass ceiling*, a social barrier that, without any form of legitimacy, obstacles minority members that are discriminated against solely based on a specific attribute. The question is if algorithms are neutral with respect to these distortions.

Empirical evidence based on Instagram social media [225, 378] prove that AI can increase the speed at which disparity grows. Once again, algorithms can profoundly shape users' context; moreover, they cannot by themselves optimize relevance and at the same time prevent the formation of a *glass ceiling*.

As mentioned before, recommendation algorithms provide users with information and people that can be added to their network because the lifeblood of social media is active users who engage with content and people [166]. To boost engagement rate, social media also recommends lists of users, which creates multiple issues.

The so-called *rich get richer effect* is one of the first identified consequences of the introduction of social media of those recommendation lists where well-known users (with high degree) are recommended proportionally to their degree so this causes a cumulative advantage which has led to hinder the emergence of content and users which are not already well-known and this could be identified as the first undesirable property of recommender. The original "**People who may know**" implemented in Twitter was based solely on topological information without any information regarding user profiles or past interaction information [166].

Plous, [316], suggested that people's opinion is influenced by the ones of their neighbour. The abundance of interactions in social media increases opportunities to create communities of like-minded people, fostered on the individual side by homophily and by AI recommender on the algorithmic side. This process can create what scholars define as echo chambers [140]: spaces where the correlation between people's opinions in a neighbourhood is strong. Empirically an echo chamber structure in social media has been addressed in multiple platforms, and results vary across platforms [86] i.e. on Facebook, a strong separation between communities with different (opposite) political leaning. This fact is not confirmed on another relevant social media, Reddit, where users' leanings are less homogeneous.

The echo chamber indirectly challenges democracies. A *democratic* decision-making process is characterized as a consensus reaching based on a majority rule. Still, the context in which each

decision is taken should be fair concerning what other people think, given the possible outcome of the various strategies that can be chosen. A people recommender system could favour the creation of an echo chamber recommending preferentially like-minded users and can skew the informational context users are exposed to and increase the influence assortment of the user's side [377]. In this setting, given the right incentives can be shown, users in echo chambers can be forced to switch their decision even in sub-optimal scenarios. Later on, Chapter 5 will show how people are prone to be misled by information personalisation because they cannot distinguish between crowd-based knowledge and peer influence.

3.2 Recommendation and potential biases

Personalized recommendation algorithms are ubiquitous on platforms independently of their scope, costs, and user profile. They suggest content and contacts, and alongside the possibility of sharing content with "friends" (the top social media system feature), they can amplify the speed and direction of information diffusion and network growth rate [383], increasing interactions between users dramatically. This poses multiple threats on different levels to discriminate between negative phenomena and a meaningful role of social media because it is not clear which is the outcome of long-term interaction (in a loop form) between the underlying AI that powers social media (which pushes content toward users) and users (that actively pull desired information).

Digital platforms can also increase the utility and create benefits for participants if they correctly manage and price network effects. The overall welfare effect of social media has been discussed in [15], and results show that both benefits and adverse effects are present. In particular, the authors point out that after four weeks of detox, people are less informed and less polarized. Subjective well-being measured with self-reported metrics shows an increase in adults [15]. In contrast, [434] demonstrate that the conceptualization of digital stress made by [375] can be applied to teenagers who, in case of problematic smartphone usage, can be affected by poor sleep quality, anxiety, and depression [434].

Serving recommendations in social media is just one step of the life cycle of a recommender system. To reach this point, each platform needs to:

1. Collect users' data and process them
2. Learn a model based on that data.
3. Provide personalized recommendations
4. Evaluate the performance

This mechanism should be placed in a loop logic because when users accept or do not a recommendation, this information will be used to refine the model. Figure 3.1 depicts the feedback loop mechanism. The basic dynamic behind the feedback loop in a recommender system is that RS collect data on the items that users view or interact with and the users' ratings or evaluations of these items, including (especially in social media) the user-user interactions. This information is then used to update the system's models of the user's preferences and interactions history, generating new recommendations. Over time, as the user provides more ratings and evaluations, the system's models of the user's preferences are refined based on previous interactions.

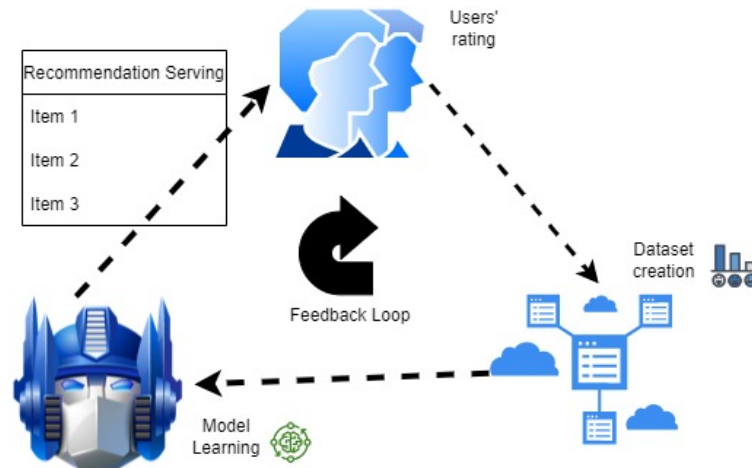


Figure 3.1: RS main component: Users, Data, Model, and the representation feedback-loop mechanism. Once the model is learned, users provide feedback that is used to refine the model.

The aforementioned loop dynamic highlights that recommender system data about users' behaviour are gathered when a recommender is already active. This creates an exposure bias because users select items based on the proposed shortlisted items, so the recommender system can be considered a confounding factor [75], in the sense that RS is a variable that impacts both the treatment assignments (which items users' view) and the outcomes (how they rate them) [234].

This can have multiple consequences, such as leading to a lack of diversity in the recommendations provided by the system and preventing the user from discovering new or unexpected items they might enjoy. Another potential drawback of these feedback loops is that they can reinforce existing biases (in the user whom recommendations are served or in the model) [229].

This can create filter bubbles where users only receive information consistent with their beliefs, escalating confirmation bias [358].

Recommendation systems can be subject to various biases that can affect the accuracy and fairness of their recommendations. Biases can arise from different sources while serving recommendations:

- **Data:** different issues can be related to data. (A) dataset shift (i.e. when the training dataset distribution of features is different from the one of the testing dataset) can affect the performances of the model; (B) popularity bias is inherent from the distribution of ratings and can generate unfair results over-recommending already popular items [56]; data are also prone to (C) selection bias given the fact that users are free to rate the items they prefer and this can inadvertently propagate bias if the users only rate items he likes (with a high rating) not providing information about what he did not like. This is also related to the problem of (D) data sparsity, meaning that most items have few or absent ratings.
- **Model:** as highlighted in Sections 2.3.3, 2.3.2 each approach has pros and cons and the specific domain and goals of the task should be included in the selection of the appropriate model.

Specifically, on the user side, the most common biases that can occur in recommender systems include:

1. Position Bias: users tend to interact with the first recommended items because, from a cognitive point of view, it is expensive to analyze and compare all the short-listed items [77, 93].
2. Confirmation bias: users will tend to select preference-consistent recommendations items, and RS can play a role in overcoming or increasing this bias [358]. For example, if an individual has previously expressed a strong preference for a particular type of item, he will tend to continue to provide feedbacks that confirm his beliefs (in this case, his preference toward a specific set of items)
3. Anchoring Bias: favoured by the fact that rating (and more in general judgements) are often influenced by elements of the environment in which this construction occurs if ratings are manipulated, displaying ratings skewed toward specific values [6, 7] the anchoring bias can influence users' rating toward that value.
4. Backfire effect: defined as a reinforcement of individual belief after a correction [389], which can lead to exacerbated polarization[30].

3.2.1 Freedom to speech and freedom to reach

Ideally, free speech is an absolute concept. In its application, it cannot be as if it is true that anyone must have the freedom to say what he wants. It is also crucial to acknowledge that words can harm others, and everyone should assume the responsibility not to harm others. In social media, this tension between freedom of speech and protecting people has been encoded into the moderation policy for content. After Twitter's delisting, Elon Musk begins its ownership indicating the direction the company will follow: "Freedom of speech doesn't mean freedom of reach. Negativity should & will get less reach than positivity"⁴ One of the main challenges in this situation is determining what constitutes harm and who has the ultimate power to make decisions about it.

It is crucial to have a transparent process in place for determining who controls the controller, as this can help to ensure that decisions are made fairly and transparently. Social media platforms are used to remove content that violates their policy, but other practices are present that can limit the spreading of content.

- **The shadowban:** or the action of restricting content distribution without user acknowledgement [250], is a current practice in social media, especially on Twitter and Instagram and can be seen as an automatic online moderation approach for preventing unwanted behaviours [227]
- **Deboosting:** it is a sort of soft ban where more actions are needed to see the content appearing (such as pressing the button "show more replies" on Twitter)
- **Search Suggestion Ban** when the search engine does not retrieve the account.

3.2.2 Information personalisation algorithms and transparency issues

As mentioned in Section 1.2, the Cambridge Analytica scandal (that involved Meta company because users' personal data had been harvested on an unprecedented scale, with more than 50 million users'

⁴Check HereMusk tweet <https://twitter.com/elonmusk/status/1598752139278532610>, retrieved on April 7, 2023

profile scraped [69]) had a significant impact on social media platforms, as it highlighted the potential misuse of user data and the lack of transparency in data collection and sharing practices. In the wake of the scandal, many social media companies have made a concerted effort to increase transparency and give users more control over their data [63]. This includes more detailed explanations of data collection practices, more robust privacy settings, and the ability to download or delete personal data. Additionally, many social media platforms have implemented stricter rules and regulations around the use of user data by third-party companies and organizations. Overall, the Cambridge Analytica scandal has helped raise awareness of the importance of data privacy and transparency on social media platforms and has led to several positive changes in this area.

The relevance of these consequences was highlighted in 2020 by a Report from Avaaz ⁵ [27] where Facebook has been publicly addressed as *danger to public health*. The report focused on dismantling a "Deception Network" of Facebook pages and accounts that spread fake news.

What is missing from platforms is a clear commitment to transparency, a value that can be effective in building users' trust [99]. Moreover, as highlighted by [64], where authors coined the term **APIcalypse**, after the Cambridge Analytical scandal, social media platforms stepped down to promote access to their data even for researchers with deleterious consequences.

Each platform indicates which specific topic is allowed or not ⁶ and the policy that they adopt to manage inappropriate content, such as the "remove, reduce, and inform" policy to counteract misinformation ⁷. Still, there is no clear and shared regulation among different platforms. The only social media platform that will probably be forced to find a compromise with legislators is TikTok, at least in the United States. As reported by many news sources ⁸ Senator Hawley (R-MO) gave birth to an initiative to review the law that ensures tech company immunity for content published by users ⁹. He proposes to maintain immunity for social media platforms under Section 230 only if they submit to an external audit that proves by clear and convincing evidence that their algorithms and content-removal practices are politically neutral. This legislation, which mainly involves TikTok, is driven by political issues such as the potential intrusion of the Chinese government on American citizen data, which led the company to switch from Singapore-based servers to the one managed by Oracle (a US-based company) ¹⁰.

To avoid a head-on collision with the European Commission, TikTok CEO plans to Meet European Union Regulators ¹¹ and in particular, Ms Vestager (the executive vice president of the EU's executive arm), whom spokesperson stated that the purpose of the meeting was to assess the company's readiness to adhere to the upcoming EU regulations about internet safety and fair

⁵Avaaz is an independent network that authored multiple investigations regarding fake news dissemination https://secure.avaaz.org/campaign/en/facebook_threat_health/

⁶Meta indicates five classes of content that cannot be recommended

⁷More here, crawled on July 7th, 2022, <https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/>

⁸The Economist reports that TikTok will let third-party audit its recommender. <https://www.economist.com/leaders/2022/07/07/whos-afraid-of-tiktok> crawled on July 7th 2022

⁹The "Ending Support for Internet Censorship Act", proposed by Senator Hawley can be found here <https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies>, crawled on July 7th 2022.

¹⁰Bloomberg highlight the fact that TikTok nears a deal with Oracle to base data-storage inside the United States. <https://www.bloomberg.com/news/articles/2022-03-10/tiktok-nears-data-storage-deal-with-oracle-for-u-s-users>, crawled on July 8th 2022

¹¹Source: <https://www.wsj.com/articles/tiktok-ceo-plans-to-meet-european-union-regulators-11673009398>, retrieved on April 7, 2023

competition among technology firms that will be implemented in 2023.

3.3 Toward a more transparent information environment

The increased power of social media comes with a need for regulation to ensure that these platforms are used responsibly and in the best interests of their users. One approach to regulation is to consider social media as a utility, similar to water or electricity, and to impose similar regulations and oversight. Another approach is to examine the role of recommender systems, which use algorithms to curate and present content to users, and to consider the potential impact of these systems on the spread of misinformation and the manipulation of public opinion. This section will explore these two approaches to regulating social media, examining the potential benefits and drawbacks and discussing the implications for policy and practice. In the end, an overview of technical solutions to improve recommendations' efficacy, including metrics beyond accuracy, is provided.

3.3.1 Are social media utilities?

A utility is a company that provides essential services to the public, such as electricity, water, gas, or telecommunications. Government agencies typically regulate these companies to ensure that they provide reliable and affordable services to consumers [433].

Several characteristics distinguish a company as a utility:

1. **Essential services:** Utilities provide services considered essential for modern life, such as electricity, water, gas, or telecommunications. These services are typically regulated to ensure that they are reliable and affordable for consumers.
2. **Government regulation:** Utilities are typically regulated by government agencies to ensure that they meet specific standards for quality and affordability. This can include setting rates, ensuring safety and reliability, and protecting consumers from abuses.
3. **Monopoly or oligopoly:** utilities are often granted a monopoly or oligopoly in their service area, meaning that they are the only or one of a few companies allowed to provide a particular service. This is often done to ensure a reliable and consistent supply of essential services.
4. **Public benefit:** utilities are expected to provide a public benefit, such as ensuring everyone has access to reliable and affordable essential services.

The main takeaway is that utilities like gas, electricity, and water are almost always distributed with a governmental intermediary who can ensure that rates are even, safety measures are being enforced, and access is as equitable as possible. Social media platforms, such as Facebook, Twitter, YouTube and TikTok, have become an integral part of modern life for many people, providing communication, connection, and entertainment. In this sense, they have many characteristics that are similar to utilities [94].

One key characteristic that social media platforms share with utilities is that they provide a sort of essential service to the public. It is also a matter of fact that Social media play a crucial role in political information and are the primary driver of news diffusion. In this sense, social media has become an essential service, much like electricity or water[24].

Another characteristic social media platforms share with utilities is that government agencies regulate them, at least partially. While social media platforms are not currently regulated as utilities, they are subject to various laws and regulations, such as privacy laws and laws against hate speech and misinformation. In this sense, social media platforms are already subject to some level of government regulation.

Finally, social media platforms also have a public benefit aspect, similar to utilities. They provide a means for people to connect and share information, facilitating the spread of essential ideas and coordinating social and political movements. In this sense, social media platforms serve a public good, much like utilities.

While social media platforms are not currently regulated as utilities, they share many characteristics with these companies, including providing essential services, government regulation, and public benefit. This has led some to argue that social media platforms should be regulated as utilities to ensure they serve the public interest.

However, there are also several arguments against considering social media platforms as utilities [396]. One concern is that such regulation could stifle innovation and creativity in the industry. Social media platforms have become successful primarily due to their ability to adapt and change quickly in response to user needs and preferences. If they were regulated as utilities, they may be required to adhere to strict rules and regulations that could inhibit their ability to innovate and adapt.

Another concern is that regulation as utilities could lead to censorship and limit free expression. If social media platforms were regulated as utilities, they might be required to enforce more strict standards for content, potentially leading to the suppression of particular viewpoints.

Finally, there are also concerns about the cost and accessibility of social media platforms if they were regulated as utilities. While most social media platforms are currently free to use, it is possible that regulation as utilities could lead to the introduction of fees or other costs. Additionally, there may be concerns about accessibility and equity if social media platforms were regulated as utilities. Some individuals may not have access to these platforms due to financial or technological barriers.

In conclusion, the question of whether or not social media platforms should be considered utilities is a complex and controversial one. While there are compelling arguments in favour of such regulation, there are also significant concerns about the potential impacts on innovation, free expression, and accessibility.

3.3.2 Recommender systems and their role as editorial board

Using personalization algorithms in social media has raised concerns about the potential effects on the diversity and reliability of the information users are exposed to. On the one hand, personalization can provide a more tailored and relevant experience for users, leading to increased engagement and satisfaction, but for example, on Twitter, mixed effects are found by [36] that concludes that algorithmically curated feeds may exacerbate partisan differences in exposure to different sources and topics.

On the other hand, it can also result in the creation of potential threats to society as highlighted in Section 3.1 where users are only presented with information that aligns with their pre-existing beliefs and biases, limiting their exposure to new ideas and perspectives or amplifying existing biases in the data that trained the machine learning models, (i.e. the skewness toward Caucasian faces in

the most face-images dataset [120, 221, 228] highlighted by [203]).

This parallel between social media personalization and the editorial process of traditional media raises essential questions about the role of technology in shaping the information landscape. Like an editorial board, the algorithms used in social media platforms make decisions about what content to promote and what to suppress, often based on complex and *not-easy-to-interpret* criteria. However, unlike editorial boards, these algorithms are less regulated¹² and consequently less transparent and are not accountable to any external authority even if algorithmic auditing has gained attention [34, 413]. Furthermore, personalization algorithms can reinforce existing power structures and inequality, leading to unfairness in the information diet provided or sub-represent minorities or entire categories (such as women) [248]. Researchers are attempting to include editorial values into the recommendation[245], and results show that it is possible to include *timely and fresh content* feature of RS without affecting the accuracy of the information provided w.r.t. users' preferences.

As technology plays a central role in shaping the information landscape, it is crucial to start acknowledging that recommender systems can act as an editorial board.

3.3.3 Improve models: going beyond accuracy

The fundamental predicament that all the stakeholders involved in social media deal with are how to leverage algorithms to benefit users in the short and long term, ensuring an experience that won't affect users or other stakeholders. This poses ethical, political and financial issues that are above the scope of this thesis, but the scenario highlights the relevance of the challenge. Usually, recommender systems are evaluated based on metrics that only consider the model's predictive power in terms of accuracy for the estimated users' preferences. To leverage positive aspects of information personalisation, a more fair approach should include the novelty and subsequent surprise that a given item recommended to a user generates. In other words, the serendipity propriety of recommender systems [188] should be included in evaluating its performances, or at least a way to include a heuristic to balance the mere accuracy of the recommender should be introduced.

Novelty and serendipity are both related to unexpectedness. Still, in the case of a novel recommendation, the users might have found the items alone, while a serendipitous system should focus on serving items that will be hard to discover. This multiplicity of aspects highlights a tension that can be summarised with the expression **exploration-exploitation dilemma** [410] where the exploitation focuses on items that best-fit users' preferences concerning users' preference. In contrast, exploration focuses on recommending randomly selected items to reduce selection and exposure biases, but this is also the baseline strategy to improve the serendipity of a recommender. The value of exploration is well-known in the context of Reinforcement Learning, but [79] tried to measure the impact of exploration using four metrics in the context of recommendation. The authors wanted to evaluate exploration's impact in the long term because exploration benefits usually do not occur in the short term, and they connect serendipity to improved long-term user experience, using as a proxy metric the conversion rate of casual users into long-term users.

More *holistic* and inclusive approaches have been proposed by scholars:

- [1] introduces **Multi-Stakeholder** approach where the quality of algorithm performances is assessed across multiple groups of stakeholders. Each one has a peculiar utility function, and

¹²RS do not have to respect the journalistic code of conduct.

recommendations should find a Pareto optimal solution before being deployed ¹³.

- [342] proposes **Collective Well-Being** as a group-level indicator measured across different dimensions that can be improved by stretching efforts on community-level key characteristics.
- [190] apply to recommendation task **Active Learning** [399], exploiting dialogue interfaces and natural language processing models to allow human-machine interaction and provide personalised movies recommendation based on "guided-sampling procedure" that only takes into account informative samples.
- Without operationalising them, in [263], authors highlight the ethical challenges of RS related to utility (in the case of inaccurate recommendations), and Rights (in cases of unfair treatment): (I) Inappropriate content, (ii) privacy, (iii) Autonomy and personal identity, (iv) Opacity, (v) Fairness, and (vi) Social effects.
- **Trustworthy Recommender system (TRec)** [125]: in this review, authors propose a multidimensional evaluation process that takes into account six dimensions: *Safety & Robustness, Non-discrimination & Fairness, Explainability, Privacy, Environmental Well-Being, and Auditability & Accountability*
- A collaboration between social media and interdisciplinary departments [381] review the process and the values to include values in RS from **Value-sensitive design**. Authors synthesize the most urgent value into *Usefulness, Well/being, Legal and Human Rights, Public discourse, Safety*

These approaches implicitly highlight that different and heterogeneous actors populate these digital platforms. Their different behaviours and needs must be considered simultaneously to ensure an equilibrium between the competing interests and goals. Each aspect in the user context (social, algorithmic, static or dynamic) should be conceptualised and operationalised to increase digital well-being.

3.4 AI and human value alignment

It should be clear that in AI, modellers do not tune each parameter manually. Their capability is only limited to defining the objective function (See Section 2.2.4) in a supervised setting and could be problematic to define an objective function that correctly takes into account all the relevant stakeholders [22]. Programmers are prone to instruct models with ambiguous or mistaken instructions. The need to encode human values into AI brings out a few but complex questions:

- Who should decide which human values should be encoded in AI? (Normative side)
- How to include those values into AI? (Technical side)

¹³Pareto-optimality is reached when, in a utilitarian framework, it is impossible to increase agent utility without decreasing someone else's utility.

As AI systems become increasingly advanced and capable of attracting investment ¹⁴, it is important to ensure they are aligned with human values to avoid potential ethical dilemmas and negative consequences. Moreover, efforts from scholars [291] and governments are undergoing to increase the accountability of AI with the objectives of increasing:

1. **Compliance:** the capability of AI to be aligned with human values
2. **Report:** indicates the degree of ability to provide or permit satisfactory explanations of their decisions [147].
3. **Oversight:** the possibility of auditing the algorithm by external authorities. This requirement is also present in the AI Act approved by the EU on 21 April 2021 ¹⁵

The first aspect concerning alignment involves defining the relevant values to humans and designing AI systems with goals consistent with these values. For example, an AI system might be designed to prioritize the well-being of humans, respect individual autonomy, or promote fairness and equality.

This involves designing AI systems that can make ethical decisions and avoid actions that would violate human values. The correct execution of this flow requires a correct ethical operationalization of the AI governance, that following [271] can be deployed and monitored using **Ethics-based auditing (EBA)** that is a structured process whereby an entity's past or present behaviour is assessed for consistency with moral principles or norms.

Ensuring value alignment in AI is challenging, as it requires a deep understanding of human values and the ability to design AI systems that can effectively uphold those values. It is also a dynamic process, as human values may change over time, and AI systems must be able to adapt to these changes.

Abel *et al.* [5] argue that Reinforcement Learning (RL) achieves the appropriate generality required to theorize about an idealized ethical artificial agent. A first step could be combining examples of social media events and corresponding simulations with automatic metric extraction algorithms, such as inverse reinforcement learning (IRL) [285]. This would allow the community to reason on more concrete examples of future conditions of the community and avoid the necessity to encode several trivial aspects that could be extracted automatically, as proposed by the Cooperative Inverse Reinforcement Learning (CIRL) framework [167].

Another approach proposed by Peschl *et al.*, [315], called *MORAL* (Multi-Objective Reinforced Active Learning). This framework combines the learned reward functions (from multiple experts) and inverse reinforcement learning. Wu *et al.*, [438], propose an approach based on the assumption that most human behaviours, regardless of which goals they are achieving, are ethical. They integrate human policy with the RL policy to achieve the target objective with less chance of violating the ethical code humans usually obey. As aforementioned in Section 3.1.5, 3.3.2, stakeholders and users can have conflicting interests, and unaware violations of users' rights can also happen, so this approach can have several limitations even if it is possible to reach Pareto-optimal solutions.

¹⁴AI state-of-the-art is also related to the capability to generate revenues, and OpenAI's recent pitch to investors said the organization expects \$200 million in revenue next year (2023) and \$1 billion by 2024 only from ChatGpt conversational model. Source Reuters <https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/>, retrieved on April 7, 2023

¹⁵Check Art. 14 here <https://artificialintelligenceact.eu/the-act/>, retrieved on April 7, 2023

3.5 Alignment and Definition of Collective Well-Being Values

As highlighted before, AI and human value alignment is a crucial area of research in artificial intelligence. What is proposed in this thesis is to align the values of AI systems with those of humans, taking advantage of a collective well-being metric defined using participatory practices and grounded on educational activities. This metric would take into account the well-being of not just individuals but also the well-being of society as a whole, including the different stakeholders [1]. Following this path, AI systems can be designed to make decisions that promote society's overall well-being rather than just maximizing individual or corporate interests. This approach can help ensure that the development and deployment of AI systems align with the values and interests of all members of society. In theory, a social planner should want to address the market failures that lead to distortions, which would take the form of increasing information about the state of the world and increasing incentives for news consumers to infer the actual state of the world. In practice, social media platforms and advertising networks have faced pressure from consumers and civil society to reduce the prevalence of fake news on their systems. Collective well-being on social media can be characterized by many different factors, including the level of trust and support among members of the group, the level of engagement and participation within the group, and the extent to which the group can achieve its goals and objectives.

Other factors that can affect collective well-being on social media include the level of diversity and inclusion within the group, the level of conflict and disagreement, and the level of support and resources available to the group from outside sources. Overall, collective well-being on social media is an important concept that can help to understand and evaluate the health and happiness of groups within a social media platform. By measuring and tracking key indicators of collective well-being, it is possible to gain insights into the factors that contribute to the success and happiness of groups on social media and to identify areas for improvement and intervention. There may be a trade-off between maximizing profits and maximizing users' well-being for social media companies.

On the one hand, social media companies primarily focus on generating profits for their shareholders. As such, they may be incentivized to prioritize activities and strategies that maximize their revenue and profit margins, even if these activities and strategies do not necessarily promote the well-being of their users.

On the other hand, the well-being of users is also an essential consideration for social media companies. If users are not happy and healthy on the platform, they may be less likely to use it regularly, which can impact the overall health and success of the platform.

Additionally, concerns about the impact of social media on users' well-being have become a more prominent issue in recent years. As highlighted by [51], participation provides a means to incorporate the broader public into the development and deployment of AI systems. ML community must change its current development practices, which are technical, representationally unbalanced, and non-inclusive, to achieve its goal of supporting people and improving prosperity.

Concerning the effectiveness of participatory practices, Arnstein [26], in 1969, used to say that "Participation without redistribution of power is just an empty and frustrating process for the powerless."

Due to the interdependence between users' behaviours and the quality of their experience on social media, to develop social media platforms that enable all community members to achieve their

desired experience without negatively impacting others, it is necessary to clearly state the values underlying the desired experience on the platform; or in other words: to define the rights and boundaries as well as the expectations that users and algorithms must respect on the platform.

All these concepts are closely related to the multifaceted and interdisciplinary topic of collective well-being (CWB).

Generalizing the concept of the individual, subjective well-being, [342] presents a structured and evidence-based characterisation of CWB for physical communities adopting several dimensions, such as vitality, connectedness and others. Ognibene *et al.*, [296], extensively discuss how to translate it to hyper-connected virtual societies like social media and the complex trade-offs between different users' needs and duties, preferences and constraints.

For example, some users must be protected from behaviours others consider harmless and enjoyable. Others would prefer being exposed to rumours instead of having to refrain from getting others' attention while sharing made-ups or avoiding missing out on some unchecked, intriguing, and maybe polarizing information.

To this, one must add the different cultural factors which would be reflected in the kind of information and interaction they would desire to have on the platforms.

While any regulation of the platform may seem better than the profit-oriented one that currently dominates them and seems to aim at trapping users in a loop of continuous consumption of content, it is easy to imagine that finding solutions for the multitude of trade-offs like the ones described above that are at least harmless for most of the highly diverse set of social media users is not a task that can be quickly and permanently addressed, asking for a continuous process that involves the users of the platform themselves.

Adopting participatory practices to design the community's collective well-being may have several advantages.

For example, involving the users in the design of the novel CWB-aimed social media community may increase their engagement in realising a desirable social media experience and, there-off, directly increase the quality of the outcome, building trustworthy and more connected communities leading to more inclusive and robust policies and therefore increasing the CWB of the community [425]. Moreover, once given the possibility, users often want to be involved in designing products or services that affect a large share of the population.

However, the co-design of a social media experience and connect CWB definition is challenging as it may require that the participating community members have an understanding of the direct effects their choices and behaviours have on other users, especially when they are integrated and potentiated through the platform algorithms. Thus a participatory approach requires an additional effort to raise users' media literacy.

This process of stating the social media community principles corresponds to defining and measuring the desirability or collective well-being (CWB) value of a specific condition of the community and is also the starting point in defining ad-hoc metrics of the overall impact of social media on the community, both at the individual and societal levels.

While an abstract and verbal description of the principle may be most beneficial for the human components of the social media community, producing a numerical metric is particularly important to direct the algorithmic mechanisms underlying the platform [136].

Participatory practices are recently gaining interest in the artificial intelligence field [51], and

their application to the definition of CWB has to overcome several issues other than the media literacy one already presented.

It is easy to imagine that the number of aspects and details to be defined would go easily beyond the operational bandwidth of a typical participatory decision process and that technologically supported approaches would be necessary.

Scholars have addressed the issue of well-being-oriented recommender systems. Khwaja *et al.*, [208], highlight that, to improve WB, daily activities recommendations should be aligned with personality traits. Leveraging the *exhibited personality*, authors build a *congruence model* that helps improve subjective well-being (SWB). From the perspective of this thesis, the main limitation of that study is that it does not take into account the collective aspect of well-being that is influenced by the social context. Alcaraz-Herrera *et al.*, [14], proposed the EvoRecSys, a framework for recommendations that uses evolutionary algorithms as its core recommendation engine. This framework models personalized well-being recommendations as a multi-objective optimization problem. The system requires as input some information about the physical status and food preferences of the users along with a well-being goal. The output of the RS is a *bundle* of activities. A user study showed that the ratings of recommendations are positive, and a challenge study highlights the superior performance with respect to a classic Collaborative filtering approach.

The next section will highlight how to build safer social media, how education plays a crucial role, and finally, an educational companion is proposed.

3.6 Educational Social Media Companion

Social media have been shown to contribute to collective well-being by enhancing people's levels of social connectivity. However, well-being, and in particular the one of teenagers, is vulnerable to social media threats, such as exposure to many types of unwanted or toxic content [98, 269].

Increasing social media users' digital literacy [127], and citizenship [196, 440] may counter most SM threats that thrive due to users' lack of awareness and over-reliance on algorithmic recommendations [37, 261, 420].

The traditional media literacy approaches were based on the idea that media adversely affected children. Therefore, it was necessary to *immunize* young people to resist such negative influence.

As the media ecosystem evolved, so did media literacy. It soon included a paradigm shift towards education and risk prevention concerning the web, video games, social networks and mobile devices. Recently, new concepts have been developed to name these new forms of literacy, from *digital literacy* or *digital citizenship* to *new media literacy* [359, 440].

With the objective of contrasting social media threats, several countries have introduced educational initiatives to increase the awareness of students with respect to the detection of fake news and misleading information on the web ¹⁶.

Still, due to their limited duration and high costs compared to purely entertaining use of social media, the effects of these programs may be limited.

The proposed framework is based on a virtual *Educational Social Media Companion* that enables continued, both in the classroom and outside, educational and interaction support for a community

¹⁶Source:<https://www.bbc.co.uk/programmes/articles/4fRwvHcfr5hYMMltFqvP6qF/help-your-students-spot-false-news> BBC, (UK), and <https://literacytrust.org.uk/programmes/news-wise/NewsWise> (UK)

of learners, creating an *Educationally Managed Social Media Community* aimed at improving users' new media literacy and social media experience. Through companion support, the students can safely *learn-by-doing* how to deal with social media content, leveraging the positive aspects and counteracting the inherent threats.

While previous educational attempts have focused on literacy activities mainly about *external* threats, improving the impact of social media on society is challenging essentially because the interactions between users determine the quality and consequences of their experience. Rising awareness about the effects of own actions on the community members' experience and the importance of performing healthy interactions to realize a desirable condition notwithstanding the anonymity [311, 356] and deindividuation that social media may foster [109, 244, 319] is central in the presented educational endeavour.

Educationally managed communities are proposed in the description of a shared vision of a “desirable social media community” in terms of an operational **Collective Well-Being (CWB)** definition specific for their community.

This will support the coherent formulation of community regulations, objectives and educational activities that involve several ethical issues entailing the definition of boundaries and trade-offs to own personal behaviour online (see Section 3.8.1), such as enabling collective satisfaction and preserving the right to free speech [426] while facing the conflicts generated by users' different attitudes, opinions, personal history, and conflicting interests.

A formalization of the CWB informs the CWB-RS, the companion recommender system aimed at recommending educational activities and content while balancing the recommendation incoming from the external social media platforms to improve the community's collective well-being, see Section 3.9.

3.7 An Educationally Managed Social Media Community

The Companion safeguards teens' interactions on social media and implements *playful adaptive educational strategies* to engage and scaffold them considering personalized *educational needs and objectives*.

These strategies comprise *scripted learning designs* [21] that, informed by the CWB-RS, will articulate the behaviour of the Companion presenting teens with the right level of educational scaffolding [44] through an adaptive, personalized and contextualized sequence of *learning activities* and supported social media interaction – incorporating behavioural and cognitive interventions (*nudges* and *boosts*) that are grounded in behavioural psychology [176, 325, 395]. Game mechanics based on a *counter-narrative* [101] approach will support learning activities related to rising awareness: motivation, perspective taking, external thinking, empathy, and responsibility.

These narrative scripts pursue collective and individual *engagement* with the Companion, offering motivating challenges and rewards to keep users' interest even in the presence of non-educational social media platforms [409] while maintaining awareness of the digital addiction threat.

The autonomous capabilities provided by the CWB-RS to the Companion can be beneficial outside of the classroom to avoid the cognitive overload, addiction or over-exposure to toxic content that the recommender system of an external, non-educational, social media platform may select. Moreover, they allow for achieving a level of availability comparable with that of non-educational



Figure 3.2: *Sketch of Companion User Interface* The Companion will support the students’ interaction with social media by contextualizing the content to increase students’ awareness and allow them to access a more diverse set of perspectives [58] and sources. It also explicitly and visually provides the students with an evaluation of the content’s harmfulness [135]. The example shows how a piece of imaginary fake news would be contextualized.

social media while reducing the moderating effort requested from the moderating educators.

Educators and the companion: a human in the loop view

In the proposed framework, the educators not only use the companion for delivering tailored educational activities in the classroom but, together with the experts, participate in the moderation and support of the community as well as in the definition of its CWB and related educational strategies, which drive the Companion by informing the CWB-RS. The educators oversee the CWB-RS behaviour playing a key “human in the loop” role [293, 448]. This alleviates the complexities faced by the CWB-RS, such as noise in the estimation of content toxicity (see section 3.1.3), which may also lead to misinterpreting users’ needs and possibly exacerbating their condition.

While the CWB-RS will have implicit moderating behaviours, e.g. reducing the presentation priority of users’ confrontational interactions, the educators will have a central role in arbitrating users’ disputes as well as solving the conflicts that may emerge between different components of an ‘under-construction’ CWB measure, such as between emotional health [342] of one user and freedom of speech of another.

Adopting Behavioural Economics to Support Collective Well-being

This educational effort aims to help users of social media make the right decision and teach them the necessary skills to get to that point. Strategies developed in the context of behavioural and cognitive sciences offer a well-founded framework to address this issue. In particular, nudging [395] and boosting [176] are considered to be two paradigms that have both been developed to minimize

risk and harm – and doing this in a way that makes use of behavioural patterns and is as unintrusive as possible.

Nudging [395] is a behavioural-public-policy approach aiming to push people towards more beneficial decisions through the “choice architecture” of people’s environment (e.g., default settings).

In the Companion context, such beneficial decisions could be to explore a broad range of different opinions about a specific topic and check understandable but scientifically correct pieces of information.

In this working example, nudges could be implemented through a visual layout of the feed that allows easy exploration of such information (see figure 3.2 below). Other forms of nudging are warning lights and information nutrition labels as they offer the potential to reduce harm and risks in web searches, e.g. [457].

The limitation of nudges is that they do not typically teach any competencies, i.e. when a nudge is removed, the user will behave as before (and not have learned anything). This is where boosts come in as an alternative approach. Boosts focus on interventions as an approach to improve people’s competence in making their own choices [176].

In the Companion context, specific educational activities have been designed aimed at teaching people skills that help them make healthy decisions, e.g. select/read/trust articles from authoritative resources rather than those reflecting (possibly extreme) individual opinions.

The critical difference between a boosting and nudging approach is that boosting assumes that people are not merely “irrational” and therefore need to be nudged towards better decisions. However, such new competencies can be acquired without too much time and effort and may be hindered by the presence of stress and other sources of reduced cognitive resources. Both approaches nicely fit into the overall approach proposed here. Nudges offer a way to push content to users, making them notice. Boosting is a particularly promising paradigm to strengthen online users’ competencies and counteract the challenges of the digital world. It also appears to be a good scenario for addressing misinformation and false information, among others. Both paradigms help us educate online users rather than imposing rules, restrictions, or suggestions on them. They have massive potential as general pathways to minimize and address harm in the modern online world [218, 242].

Educational Activities

The Companion must also provide a satisfying and engaging experience by using *novel hand-defined educational games and activities* based on the interactive counter-narrative concept and educational games. Social media’s entertainment aspect is preserved during the navigation modulated in taking into account CWB, suggesting activities, content, and contacts for the user but managing the exposure to potential threats and addiction.

The Narrative Scripts help raise users’ awareness about SM threats and train the students against them. They are sequences of adaptive learning tasks that provide the right level of educational scaffolding to individuals in developing critical thinking skills, including awareness, perspective taking, motivation, external thinking, empathy, and responsibility by interacting with narratives, counter-narratives, and peers. These tasks can be different activities, including free-roaming inside the platform, guided roaming following a narrative, quizzes, playing minigames, or participating in group tasks. Different counter-narratives can be triggered depending on students’ detected behaviour[237].

Counter-narratives are used to challenge biased content and discrimination, highlight toxic aspects of messages and attitudes, challenge their assumptions, uncover limits and fallacies, and dismantle associated conspiracy and pseudo-science theories.

Through a game-oriented setup, the companion bridges the “us” versus “them” gap that is fostered by hate speech and other expressions of bias (e.g., gendered) and brings forward the positive aspects of an open society and focuses more on “*what we are for*” and less on “*what we are against*”. The users will be informed and requested to actively and socially contribute to creating and sharing content and material that fosters and supports the idea of an open, unbiased and tolerant society. Thus, the games can also offer the chance to build connections between the users, who are more vulnerable to toxic online content when isolated. One approach is periodically proposing specific tests and activities related to each threat, such as [390].

3.8 Defining a Collective Well-Being Metric for Social Media

Social media is an integral part of our everyday lives that is having both negative and positive effects [78, 421]. Hence, as positive aspects rely on the same mechanisms exploited by threats, and because each user’s behaviour will affect the other members of the community while values can differ between communities, it is desirable and necessary to explicitly and collaboratively define shared community principles corresponding to the desired condition of the community. These community principles will constitute the foundation to define a specific measure of the overall impact of social media in the community at an individual and a societal level, that is, to measure the desirability or *Collective Well-Being (CWB)* of a certain condition of the social media community [342]. These community principles, formalised in the CWB measure, together with an understanding of the virtual and physical social dynamics in the community, should drive the definition of users’ behaviour guidelines and connected educational objectives to reach and maintain the community in the desired condition, or in other words, to achieve a high level of CWB. A quantitative measure of CWB allows for a more accurate evaluation of the impact of different aspects of the interaction on the community while taking into account the complex and fast dynamics of social media. When CWB is estimated directly on the SM platform it could allow directing its autonomous components, e.g. recommenders, to collaborate in achieving the desired community condition. This would be a more democratic and transparent objective than the ones currently pursued by the social media platforms [155]. The proposed framework is used to direct the algorithms at the interface between the educationally managed community and external social media.

3.8.1 Research on collective well-being and social media

The literature presents several definitions and measures of well-being [142, 400]. Some of them were applied in the context of social media to estimate their effects [78, 220, 268, 415, 421] but mostly considering the single individual with limited consideration for the overarching social aspects [174].

Gross Domestic Product (GDP) has been proposed as an index of the economic well-being of a community ¹⁷. In such contexts, inequality is also an important factor, and it is common practice to use the Gini index to measure it [300]. While the economics view is difficult to connect to a social

¹⁷Retrieved from: <https://voxeu.org/article/defence-gdp-measure-wellbeing>

media context, they share similar key issues: which aspects to measure and, above of all, how to compare and aggregate measures of individuals' well-being to synthesize that of the whole society [97], even if in this work we consider only the local educational community.

Multidisciplinary notions of CWB extend that of individual well-being to measure a group-level property (construct). They include community members' individual well-being incorporating diverse domains, such as physical and mental health, often stressing the presence of positive conditions. They study which properties of the community affect the members and how much each of these properties adds to a comprehensive measure of collective well-being. Have been already stressed the importance of education and educational objectives to support constructive interactions and achieve desirable community conditions, i.e. a high level of well-being. However, education itself is often already part of well-being frameworks [262, 342, 372, 429]. The connection between education and well-being has been analysed from several perspectives. In this framework, the most relevant one is the one defined as *social and emotional literacy* in [371].

Roy *et al.*, [342], present a CWB framework divided into different domains and comprising health-care and non-health-care-related community factors where the contribution of the latter ones is supported by evidence of their effects on health. This framework can help to define a checklist for the definition of a community-specific CWB and related measures and indicators:

- *Opportunity* domain is related to "the perceived opportunity to achieve life goals and socio-economic mobility" [110] as well as the access to education. Social media can be a powerful tool for accessing many opportunities. Feeling in control while using them, instead of just a distraction or worse an addiction, may be an important part of CWB for SM;
- *Connectedness* domain is related to the presence of supportive, high-quality, reciprocal relationships with secure attachments. Includes dimensions of social acceptance and social integration that depend on the behaviour of other members of the community [419];
- *Vitality* domain covers many emotional aspects of several individual well-being definitions, such as Fredrickson's one and Seligman's model of flourishing [132, 360]. However, spillover effects [174] and emotional influence make vitality an important aspect also at a social level; The threats presented in Section 3.1 would impact negatively the affects component of the *Vitality* and *Connectedness* domains;
- The *Contribution* domain relates to community engagement and related feelings of meaning and purpose. Contribution can improve other members' experience but may also have negative effects;
- The *Inspiration* domain relates to creativity and lifelong learning, areas where social media have a huge potential.
- The psychosocial *Community* characteristic that is relevant for social media settings:

"A community with a negative psychosocial environment is segregated and has high levels of perceived discrimination and crime, high levels of social isolation and low community engagement, and low levels of trust in government and fellow citizens."
[118, 215, 246].

Community is partially overlapping with the Connectedness and Contribution domains but describes aspects that are easier to concretely measure in social media networks.

While these formulations of CWB can inspire a guideline to define social media communities' principles and CWB metrics, they must be extended and formalized to better take into account the specific issues and opportunities of SM and in particular, the threats reported in section 3.1. Another important aspect to address is combining contrasting factors or, in other words, formalizing the complex ethical decisions induced by the conflicts and trade-offs that emerge in any social context [276].

Challenges of defining collective well-being for social media

Defining a CWB metric for SM is an ambitious endeavour that requires a combined effort of different disciplines. It would range from political sciences, sociology and psychology over ethical considerations to computer science, machine learning and network theory.

Besides CWB aspects for physical societies, the impact of integrated intelligent agents must also be taken into account in the context of social media, as discussed in section 3.1.5.

A CWB measure for virtual communities has to take into account the conflicts between members as they are frequent and algorithmically augmented. Therefore, the conflict between the right to freedom of expression, user satisfaction, and social impact must be stressed more when defining a social media CWB than with physical societies where these factors have slower and better-understood effects and may have regulations already in place [426].

Conflicts between members' interests pose serious ethical concerns that are out of the scope of this thesis and have been the focus of recent research in AI and ethics in different domains [73, 211, 264]. When social media are integrated into an educational framework, the problem may be mitigated by involving educators and experts as moderators. Such an educational setup can also allow initial studies of the implications of a social media platform that aims to improve CWB.

3.8.2 Participative definition of social media community principles and CWB factors

Social media community principles and corresponding CWB factors must be shared by the members of the community. While research in the field can inform about common social aspects, internationally acknowledged human rights, or social media-specific phenomena, a community would most likely have the freedom to define tailored principles. To achieve this human-centred approaches to the participatory design of technology are being explored by researchers. These approaches involve the stakeholders in the analysis of relevant factors and the co-design of technological solutions. One of the main challenges is bridging the gap between the community members' knowledge and the complexity of cyber-social systems like social media [105]. An example is a qualitative study to explore adolescents' representations of social media based on pictorial metaphors, reported in [391]. The study proposed and analyzed the outcomes of a school project entitled "The Social Media of the Future". Discourses and visual representations of a total of 168 drawings about their visions for their ideal Social Media tools were analyzed. The results of the analysis pointed out that the relevant CWB factors shared by the adolescents participating in the study were: care about additive

features, transparency in the conflict of interest behind the SM business, also in terms of agency to be able to monitor and control privacy and security facets.

3.8.3 Toward the automatic estimation of collective well-being in social media communities

Social media are strongly integrated with information systems that can affordably offer a huge amount of data with a high frequency. Transforming this data for the estimation of suitable collective well-being measures through machine learning methodologies would open the way to many research and applicative opportunities, such as autonomous systems that maximise CWB and avoid current issues induced by profit-based objectives.

Current CWB formulations are not easy to estimate directly using data available in real-time on social media, which is necessary to support an autonomous system optimizing CWB. Moreover, such formulations need to be extended to take into account specific social media issues. For example, most of the available formulations of collective well-being focus on positive aspects. Nevertheless, the positive aspects (see sec. 3.8.1 and negative ones (see sec.3.1) need to be explicitly considered as part of the CWB as they strongly affect social media users and in particular teenagers.

In this thesis, a definition of *collective well-being metric for social media* is proposed by combining suitable components of classical CWB and SM threat measures. The measures of these components could be measured by periodically proposing specific surveys and activities [243]. However, additional richer and more transparent measurements can be performed by developing intelligent components that analyze users' behaviours.

In this definition, for each user, event, i.e. content or connection related, and aspect defined relevant for the CWB three terms are computed:

- **CS(aspect, user)** Content Shared measures the aspect-specific value of the content shared by the user;
- **CE(aspect, user)** Content Exposure measures the aspect-specific value of the content observed by the user;
- **CC(aspect, user a, user b)** Contact Creation measures the aspect-specific value of new connections based on the participants' CS and CE.

These elements account for the double role of each member of the social media community as both receivers and producers of content. In the proposed educational setup, where only the community of interest is in contact with an external social media community, it is distinguished between "*endogenous*" and "*exogenous*" aspects. The community can be exposed to threats that are generated outside but a community can also generate such threats inside as part of the interactions in the social medium. In this case, the feeds from external sources may be weighted differently.

While the CS can be seen as a direct expression of the state of the user, it strongly depends on the user's style of interaction. Moreover, only relying on the content shared by users would induce a substantial delay compared to the moment when a user got affected by observing a piece of content (CE).

Conversely, the user is exposed to a multitude of diverse inputs hindering the interpretation of the overall effect only from the CE, while the user's reactions (CS) may be more indicative of

the most impacting events. Indeed, current affective state estimators and toxic/positive content detectors can only provide noisy estimations of the current user state and the content quality. However, the availability of complementary data with higher reliability is limited.

Once each event is scored for each aspect of interest, it must be decided how to aggregate these terms over users, time, and the different aspects to obtain an estimation of the total CWB of the community. Indeed, the definition of an actual metric following this strategy requires making several choices. For example, about the scale for the terms of different aspects considered. Regarding aggregation over time, CC, CE and CS values could be simply averaged. Other approaches could be considered to take into account the frequency of the events or the diversity of opinions presented or give more relevance to extreme events, which may be more accurately detected and evaluated. In particular, the value of being exposed to multiple opinions (time-aggregated CE) may be augmented with a measure of diversity (e.g. entropy) [139, 253].

The design of the CWB metric presents several challenges requiring careful consideration even for small educational communities that the proposed framework targets. In devising their solutions often the naive approach may at best be ineffective, and at worst exacerbate the issues it was intended to solve. For example, the aggregation over the aspects dimension may not seem complex when considering the aspects to be independent. In reality, the impact of the various aspects on the users may be interlinked, for example, over-exposure to content focused on one aspect (e.g. videogames) may lead to overuse of the platform or tire the user who will lose the opportunity to learn about more important content (e.g. social issues).

The most complex aggregation to design is over users because it has to balance the well-being of different individuals and groups of users taking into account their conflicting interactions along different dimensions. It is important to consider the different features of each user while respecting privacy constraints.

For example, vulnerable users are often victims of toxic content but also producers [47, 62, 249], which affects the CS value. It is important that they are not isolated [67] and that, at the same time, the toxic content should not be fed to those who could be more affected and instead presented to educators or other community members that have shown constructive reactions to such type of content. This means that the content exposure (CE) should be differently weighted for different community members based on their resilience and that supportive connection creation (CC) should be favoured between people with high and lower resilience. Still, resilient members mustn't be overloaded with toxic content and support responsibilities [376].

Apart from weighting issues, another important formal issue is the selection of the actual aggregation function across users. Adopting the naive average a society where a few radicalized users share extremely hateful content may have a higher CWB score than one with many users sharing content about action movies with slightly violent scenes.

Another reason why a linear combination of components may not be suitable in the definition of a well-being measure is that it will simply induce maximizing the terms with positive weights and minimizing terms with negative ones, without allowing a balance. For example, if interactions between drastically opposite opinions are considered negative because of possible backfire effects and flames [30], and interactions between excessively similar opinions are also considered negative because of the echo chambers they may give place, then also interactions between moderately different opinions will have a negative value even when they may lead to a reduced polarization.

Other aggregation functions may be chosen but it is still difficult to find general solutions. For example, defining the well-being of society as the well-being of the member with lower well-being (i.e. minimum instead of average) could lead to focusing all the resources on factors that may not be changed.

3.9 An educational Collective Well-Being Recommender System

Recommendation systems (RSs) are ubiquitous in online activities and are crucial for interacting with the endless sea of information that the Internet and social media produce today. In social media platforms, they have introduced the possibility of personalizing suggestions of both content and connections based on the use of user profiles containing also social features [80, 112, 173]. Their goal has been to maximize the users' engagement in activities that support the platform itself. However, these self-referential objectives fail to consider repercussions on users and society, such as digital addiction [18], filter bubbles [58], disinformation wildfire [426], polarization [332], fairness [2, 331], and other issues discussed in Section 3.1.

To address this, the concept of *Collective Well-Being aware Recommender Systems (CWB-RS)* is proposed. The CWB-RS extends social media RS intending to maximize the cumulative long-term *CWB metric* instead of self-referential platform objectives. Compared to previous efforts in dealing with possible negative effects of RSs [2, 331, 332], the CWBRS takes into account multiple issues and, to reduce their cumulative impact on society, it adopts longer terms strategies fitting into the proposed educational framework. Integrating educational objectives aimed at achieving *CWB* in the longer term the CWB-RS will also have functions similar to those of a (collective) *Intelligent Tutoring System* [158].

RSs have been widely used in educational settings [247], and they are receiving increasing attention due also to the fast growth of Massive Open Online Course (MOOC) [339] and the availability of big data in education [361]. In educational contexts, recommendations are sequential and functional to achieving learning goals [394]. Similarly to the social media context, they have also employed social information [114, 216]. However, they are usually acting on the content provided by educators with educational aims, while CWB-RS also has to redirect disparate content flowing from external Social Media toward achieving educational objectives.

The CWB-RS creates new recommendations presented through the Companion by processing both the content generated *internally* by the members of the *educationally managed social media* community and the content recommended for them by the RSs of the *external* platform. *Content Analyzers and Threat Detectors* (see Figure 3.3 and sec 3.1.3) will analyze each piece of content to evaluate the level of threat and other relevant information for the CWB metric, such as the users' opinions and emotions (see sec. 3.8.3). This information will be used to:

1. evaluate the current condition of the users;
2. *augment and contextualize* the content provided to the users;
3. *evaluate* the future effects of different sequences of content re-rankings and recommendations through predictive models of users' conditions;
4. *select* the actions that account for the highest expected, long-term, cumulative CWB metric.

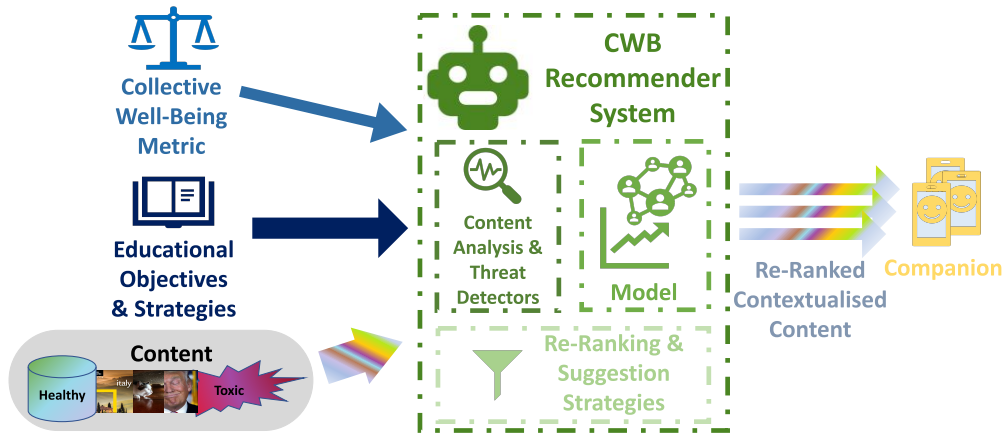


Figure 3.3: *Role of the CWB-RS in the Companion.* CWB-RS will process the *content* generated by the users of the educationally managed social media and the *content* externally recommended for them by the RSs of the external social media platform to create new recommendations aimed at maximizing the cumulative long-term *collective well-being metric*. *Content Analyzers and Threat Detectors* will analyze and evaluate the level of threat for each piece of content and other relevant information as the users' emotional state. This information will be used to (1) *augment* the information provided to the users by the companion interface; (2) *evaluate* through *predictive models of users' opinions and reactions* the future effects of different sequences of re-ranking and recommending actions; (3) *select* the re-ranking and recommending actions that resulted in the highest expected cumulative improvement in terms of learning objectives, CWB metrics, agreement with selected educational strategies and user engagement.

3.10 Conclusion

In this chapter, the social media and recommender systems potential threats have been introduced. The topic of AI-human value alignment has been addressed, and the concept of an Educationally managed social media community is proposed. The result was the development of Collective Well-Being (CWB) for Social Media communities, which refers to the shared perception of what constitutes desirable conditions for a specific community, shaping its members' educational goals and expected behaviours. With respect to other approaches, such as the *Multi-Stakeholder recommender systems*, the proposed approach is more flexible because it is *"local"* in the sense that each community can theoretically implement the CWBRS for specific needs. It also allows the deployment of educational activities and counternarrative scripts. The mentioned approach shares the limitation of defining values because, in the multi-stakeholders, it is not so easy to deploy the different utility functions for each stakeholder.

The proposed companion will allow the smooth passage from everyday use of social media to an educational experience by interfacing with the students to support and guide their interaction with the social media environment both inside and outside the classroom. In social media communities, as in any society, the safety and well-being of its members are determined by their mutual interactions.

Therefore, an important endeavour is to increase users' awareness of the consequences of their actions and acceptance of necessary boundaries, especially in such deindividuating environments.

Chapter 4

Exploring the Interplay between Social Media and Recommender Systems through Simulation

This chapter will address the role of recommender systems (for both content and other users' recommendations) concerning the potential harmfulness of such algorithms. The first section presents an introduction to these tasks, covering the components of Agent-based and opinion models. Later on, a model and a simulation procedure to mimic stylized opinion and social dynamics are introduced, helpful to understand the dynamics of opinions in groups of people who interact with each other in the context of an overabundance of information that cannot be processed entirely by any users in the network and needs to be filtered ranked in a personalised manner. The evaluation will be performed using multiple metrics that can improve the analysis or the recommender system performance with respect to the diversity and novelty of recommended content but also in relation to the homogeneity of recommended people and the impact of different strategies on polarisation in terms of the capability to create filter bubbles and echo chambers. Even if researchers have found a limited role of echo chambers and filter bubbles in increasing polarization [401], a more constructive approach than questioning whether social media is causing polarization would be to explore the potential of social media interventions to alleviate it, using machine learning for example in recommenders systems applications.

4.1 Agent-based models as Mathematical Representations of dynamic society

Even if, for some topics, opinions can be formed immediately and do not change over time, interactions and social influence affect in most real-life events the formation and the dynamics of own opinion. This complex and rich phenomenon mixes personal beliefs and interactions between individuals. Opinion dynamics has gained much attention from scholars for over 60 years with different approaches from psychology and physics to economics and computational sociology. The seminal work of DeGroot, [103], opens the path to the consensus models and is one of the first mathematical representations of opinion dynamics.

These models can generally be understood as an instance of Agent-Based modelling (ABM).

ABMs are computational models with different components [55]:

- Environment in which agents operate
- Agents that are decision-making entities
- Rules that govern interactions between agents

The foundation of network simulations is to specify the behavioural rules of each agent (each one of them can have specific attributes that characterize their current states), as well as the rules of their interaction, to simulate thousands of them using a computer model, and to explore the consequences and the outcome on the global level of the population as a whole, using results of simulation runs. Typically, ABM models are discrete models. In each run of the ABM, given a set of initial conditions and rules (that govern each agent's interactions and environment), the model is executed for a given number of time steps.

4.1.1 Advantages of Agent-based models

There are multiple advantages of using ABM models, which span from having different types of agents (heterogeneous populations) and specific rules for each type to the possibility of studying emerging phenomena. Interaction between agents can be regulated, and the micro and macro outcomes can be analyzed [143]. The variables (one or multiple) of interest are usually systematic and tend to appear on the macro-level [129]. By doing so, ABMs can capture the complex and dynamic interactions that occur within a social network. By simulating the behaviour of individual users (i.e., the agent) within a social network, an agent-based model can help researchers better understand how different factors and variables that influence the spread of information on social media. For example, an agent-based model could be used to study how the structure of a social network (e.g., the number and type of connections between users) affects the spread of information on the platform [86]. By running the model multiple times with different assumptions and inputs, researchers can explore how different factors interplay and gain valuable insights into the underlying dynamics of the platform.

4.1.2 Disadvantages of Agent-based models

One of the drawbacks of ABM is that these models' computational complexity increases exponentially with the number of actions required in each step of the simulation for each agent. Another aspect involves the model itself: to create an accurate model, researchers must carefully specify the rules and constraints that govern the behaviour of the agents (in this particular case, how opinions are formed and how they change over time and, above all, which factors contribute to this change), as well as the initial conditions of the system. This can be a complex and time-consuming process, requiring a deep understanding of the system being studied. Finally, agent-based models are only as good as the assumptions and inputs used to create them. The model may not produce reliable or meaningful results if the assumptions or inputs are inaccurate or unrealistic. The peculiar trade-off in ABM is that researchers must find an acceptable compromise between realism (high complexity) and simplicity on the computational side but also in interpreting results [387].

4.2 Agents and Opinion Spaces

Agents are the fundamental block of ABM, where interactions between them are simulated as a consequence of rules that the model designer chose. In social media, participants are constantly exposed to new information, and the interactions may cause a dynamic change of opinions. In the last decade, this process and the changes in the (increasing) polarisation of the public sphere have interested different scholars. In more formal terms, opinions can be represented differently depending on the researcher's specific needs and the topic, using discrete and continuous values, with one or multiple dimensions or numerical rating systems to measure people's opinions on a particular topic [290]. Another way to represent opinion mathematically is to use probability distributions. This approach involves assigning a probability to each possible opinion on a particular topic and then using these probabilities to calculate the likelihood that a given person will hold a particular opinion.

4.2.1 Binary Opinion

The binary opinion model is often used to represent situations where people hold one of two possible opinions on a particular topic, such as *"yes"* or *"no"*, *"agree"* or *"disagree"*, or *"support"* or *"oppose"*. A straightforward example is the vote regarding the adoption of a law by the Senate's representatives that could be represented as a binary opinion (top left of Figure 4.1). An example of ABM with binary opinions and interaction with recommender systems can be found in [314], which found that different strategies and initial opinions' distribution can promote the creation of echo chambers.

4.2.2 Continuous 1D Opinion

If the degree of (dis)agreement is a better representation of the agent's decision, as it can happen for topics such as taxation or immigration, a continuous one-dimensional opinion model (top right of Figure 4.1) is more suitable. It represents opinions on a continuous (i.e., non-discrete) scale. This model is often used to represent situations where people hold a wide range of possible opinions on a particular topic rather than just two discrete values (as in a binary opinion model). In a continuous one-dimensional opinion model, the range of possible opinions is represented by a continuous line or curve, with the positions along the line or curve corresponding to different opinions. For example, a line that ranges from -1 to $+1$ could be used to represent the full range of possible opinions on a particular topic, with -1 representing the most negative opinion and $+1$ representing the most favourable opinion. One of the key advantages of a continuous one-dimensional opinion model is that it allows for a more precise and nuanced representation of opinion than a binary opinion model.

4.2.3 Continuous 2D Opinion

A two-dimensional opinion model is a mathematical model used to represent opinion on a two-dimensional (i.e., two-valued) scale. This model often represents situations where people hold opinions on a set of topics. In a two-dimensional opinion model, the two dimensions (i.e., the two values) are typically chosen to represent two different but related aspects of a person's opinion. For example, a two-dimensional opinion model might use the dimensions of *"agreement"* and *"intensity"* to represent people's opinions on a particular topic. In this case, the first dimension (agreement) would

represent whether a person agrees or disagrees with a particular statement or position. In contrast, the second dimension (intensity) would represent how strongly they feel about that position.

Using two dimensions, it is possible to capture multiple aspects of a person's opinion and see how they relate to each other.

In Figure 4.1 the different opinion spaces are depicted

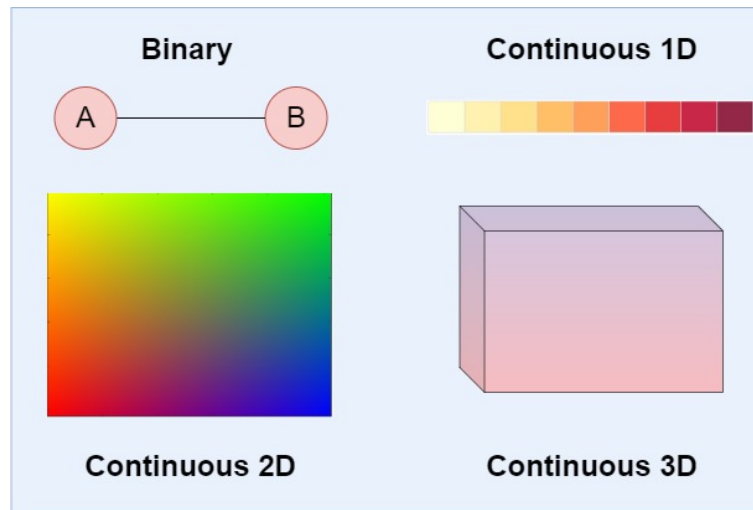


Figure 4.1: Opinion spaces: binary opinion (top right), Binary continuous opinion (top left), 2D continuous opinion (bottom right), 3D opinion (bottom left).

4.3 Society as a graph

To resemble society in the ABM model, a way of representing the relationships and connections among individuals or groups within that society is needed. This can help analyze patterns and trends within the social network to understand how certain phenomena, such as homophily, can influence the structure and dynamics of the network and to model later emerging phenomena, such as filter bubbles and echo chambers or polarization. In other words, the focus is twofold: model the individuals' dynamic but also the influence of social context on individual behaviour [200]

Usually, the starting point of modelling social networks is graph theory. Below the main building blocks are covered.

4.3.1 Graph theory elements needed for Agent-based models

Graph theory is a branch of mathematics that studies the properties and applications of graphs, which are mathematical structures used to represent pairwise relationships between objects. A graph consists of vertices (or nodes) representing the objects and edges representing the relationships between the objects.

In computer science, graphs are used to model networks, such as the internet or social media, and to solve problems related to network connectivity and optimization, such as ranking the give Google the leadership of search engines market given their proprietary algorithm Page-Rank [303]. In social sciences, graphs are used to model social networks and to analyze patterns, and trends in the network, such as the presence of communities or the influence of homophily [92].

Different graphs exist depending on the nature of the relationships between the objects. A simple graph is a graph that does not contain any loops (edges that connect a vertex to itself) or multiple edges (more than one edge between the same pair of vertices). A directed graph, also known as a digraph, is a graph in which the edges have a direction, and the relationships between the vertices are asymmetric. An undirected graph is a graph in which the edges do not have a direction, and the relationships between the vertices are symmetric.

Modelling society as a graph can involve representing individuals or groups as nodes and the connections between them as edges. Different types of users can coexist with different rules that determine their behaviours. The weight of the edges can indicate the strength of the connections, and the direction of the edges can indicate the direction of the connections. Opinions that characterise users can be included in the graph as an attribute of nodes and the same can be done for edges.

4.4 Modelling Interactions

An opinion model is a mathematical model used to represent people's opinions on a given topic and their dynamic change over time, assuming that opinions are socially influenced. Models are often used in computational social science, where researchers use them to study how people's opinions change over time and how different factors can influence those changes.

One common way to represent an opinion is with a numerical value on a scale, such as a scale of 1 to 10. This allows researchers to compare opinions and see how they change over time. For example, suppose a group of people is asked to rate their satisfaction with a particular product on a scale of 1 to 10. In that case, the average rating can be calculated and used to see how people's opinions change over time.

Opinion modelling can be a powerful tool for understanding how people's opinions are formed and how they change. For example, researchers can use opinion models to study the effects of advertising on people's opinions or to see how the opinions of others influence people's opinions. This chapter will focus on the interactions between users and recommender systems (RS) that can be intended as a super-user, influencing, under certain conditions, users and their opinion. In Figure 2.2, the main components of modelling society as a graph are reported: nodes, edges and nodes' attributes.

4.4.1 DeGroot Model

The DeGroot model (DG) [103] is one of the first mathematical models created to study how people's opinions change over time, and in particular, under which conditions it is possible to reach a consensus. It was developed by the American mathematician Morris DeGroot in the 1970s and is based on the idea that the opinions of others influence people's opinions. In the DeGroot model, each person is represented by a node in a social network. The connections between nodes could represent the existence of a relationship and/or the influence one person has on another if the edges connecting the underlying network have weights. The model assumes that the opinions of those around them influence nodes' opinions. This influence is equal for all nodes in the absence of weight or proportional to the strength of the connections between nodes if the edge's weights are present.

One of the critical features of the DeGroot model is that it allows for the influence of outside factors (external to the user model) on people's opinions. For example, suppose a group of people

is exposed to a particular information to mimic the intervention of media such as television. In that case, the model can be used to predict how that information will influence their opinion, simulating a sort of media or government that can spread the news around all nodes.

The formula for updating opinion in the DeGroot model is given:

$$O_{i,t+1} = \frac{\sum_j w_{i,j} \cdot O_{j,t}}{\sum_j w_{i,j}} \quad (4.1)$$

Where $O_{i,t}$ is the opinion of individual i at time t , $O_{j,t}$ is the opinion of individual j at time t , and $w_{i,j}$ is a positive constant that represents the weight or influence of individual j on individual i .

The DG model has been extended to include also repulsive behaviour such as backfire by Dandekar *et al.*, [100]. The goal is to model biased assimilation that corresponds to the fact when presented with mixed or inconclusive evidence, individuals draw undue support for their initial position, thereby arriving at a more extreme opinion. Findings show that homophily alone, without biased assimilation, is insufficient to polarize society. When the network is homophilous on the other side, the opinion formation process generates polarization.

4.4.2 Bounded confidence model

The family of *bounded confidence models*, [102], is a mathematical model used in sociology to explain how individuals in a social network interact. The model assumes that individuals have a certain level of confidence in their beliefs and that a certain threshold binds this confidence. This means that individuals are only willing to interact with others who share similar beliefs and will not engage in discussions with those who hold beliefs that are too different from their own. In other words, people can only understand and trust others to a certain extent, and beyond that point, their ability to do so begins to break down, and they are unwilling to interact.

The bounded confidence model can explain various social phenomena, such as the difficulty of achieving consensus within a group. For example, suppose individuals within a social network have a high confidence in their beliefs (and consequently a low threshold). In that case, they may be less willing to consider alternative viewpoints, resulting in the formation of echo chambers. On the other hand, if individuals have low confidence in their beliefs (so a high threshold), they may be more open to engaging with others who hold different beliefs. This can also make it more challenging to achieve consensus within the group. In particular, these models show how people with similar beliefs and attitudes tend to cluster together and form social groups. In contrast, those with different beliefs and attitudes may be more likely to remain isolated [143].

The formula, adapted from Hegselmann *et al.* (HK), [172], for updating opinion in the bounded confidence model is given:

$$O_{i,t+1} = \begin{cases} O_{i,t} + \mu \cdot (O_{j,t} - O_{i,t}) & \text{if } |O_{i,t} - O_{j,t}| \leq \epsilon \\ O_{i,t} & \text{otherwise} \end{cases} \quad (4.2)$$

Where $O_{i,t}$ is the opinion of individual i at time t , $O_{j,t}$ is the opinion of individual j at time t , μ is a positive constant that determines the rate of opinion change, and ϵ is a positive constant that represents the threshold of confidence above which individuals will not engage with others who

hold different beliefs.

One way to improve the bounded confidence model is to incorporate additional information transmission and opinion formation mechanisms. For example, the model could be adapted to look like a social media environment where an information personalisation system filters the opinions users are exposed to.

To be more realistic, an opinion model should also consider that opinion detection from social media algorithms could be noised. The impact of noise in the model (that could be intended as external sources of information that influence the agent or misclassification of his opinion by the information filtering system) has been investigated by [384] that provides a theoretical framework and conclude that random noise could almost induce the HK dynamics to achieve "*consensus*".

4.4.3 The Friedkin–Johnsen model

The Friedkin–Johnsen (FJ) [133] assumes that individuals have an initial opinion on a given topic and that others in the network influence this opinion. The formula for updating opinion in the Friedkin–Johnsen model is given:

$$O_{i,t+1} = \alpha \cdot O_{i,t} + (1 - \alpha) \cdot \frac{\sum_j w_{i,j} \cdot O_{j,t}}{\sum_j w_{i,j}} \quad (4.3)$$

Where $O_{i,t}$ is the opinion of individual i at time t , $O_{j,t}$ is the opinion of individual j at time t , $w_{i,j}$ is a positive constant that represents the weight or influence of individual j on individual i , and α is a positive constant that represents the degree to which an individual's initial opinion is retained when updating their opinion.

The Friedkin–Johnsen model is similar to the DeGroot model in that it also uses a weighted average to update individuals' opinions. However, the Friedkin–Johnsen model also incorporates the effect of individuals' initial opinions, allowing it to capture better the dynamics of opinion formation and change within a social network limiting the variability of opinions. The FJ's α parameter could be interpreted as the degree of users' opinion stubbornness.

Include repulsive behaviours into FJ model

The FJ model can be adapted to include repulsive behaviours by modifying the formula for updating the opinions' update process to account for the influence of individuals who hold beliefs opposite to those of a given individual. This can be done by introducing a negative weight for individuals who hold opposing beliefs, which will cause their opinions to have a repulsive effect on the opinion of the given individual.

The modified formula for updating opinion in the Friedkin–Johnsen model with repulsive behaviours is given by:

$$O_{i,t+1} = \alpha \cdot O_{i,t} + (1 - \alpha) \cdot \frac{\sum_j w_{i,j} \cdot O_{j,t}}{\sum_j |w_{i,j}|} \quad (4.4)$$

Where $O_{i,t}$ is the opinion of individual i at time t , $O_{j,t}$ is the opinion of individual j at time t , $w_{i,j}$ is a positive or negative constant that represents the weight or influence of individual j on individual i , and α is a positive constant that represents the degree to which an individual's initial opinion is retained when updating their opinion.

In this modified version of the Friedkin–Johnsen model, individuals who hold beliefs that are opposite to those of a given individual will have a negative weight, which will cause their opinions to have a repulsive effect on the opinion of the given individual. This can help to capture the influence of repulsive behaviours on the dynamics of opinion formation and change within a social network.

Sirbu *et al.*, [368], extend the bounded confidence model so that the probability of interaction between two agents whose opinions lie within the confidence radius depends on the difference of their opinions. The smaller the difference, the greater the probability of interaction between randomly chosen pairs. Findings show an increased tendency towards opinion fragmentation, an increased polarisation of opinions, and a dramatic slowing down of the speed at which the convergence is reached.

4.5 Recommender systems in simulation settings

Information filtering systems could be included in the ABM framework to simulate their influence on users' opinions and echo chamber formation. There are different ways to include a recommender into ABM and opinions model:

- To model a social media where news outlets are part of the system, Quattrociocchi *et al.*, [327], introduce a novel model that accounts for both the coexistence of media and social influence as two separated but interdependent processes.
- Perra, [314] introduces the recommender systems as an intermediate step that filters information flows between users. In his binary opinion model, the RS is introduced as a "super-node", and different recommendation strategies are tested. Results found that algorithmic filtering might influence opinions' share and distribution, especially in case information is biased towards the current opinion of each user.
- Blex *et al.*, [52], models the role of people recommender. The authors conclude that any level of positive algorithmic bias in the form of rewiring can prevent fragmentation and its effect on reducing the fragmentation speed is negligible.
- Geschke *et al.*, [143], model both the content and people recommendation system to address the issue of echo chamber formation. They conclude that even without homophily content and people's recommenders can favour the formation of echo chambers. A limitation of the model is that they do not apply state-of-the-art recommendation techniques in the literature but take advantage of a random selection of content and node and use similarity metrics to favour or counteract users' opinions. Moreover, they assume *perfect knowledge* of the system and consequently do not apply any noise to the opinions.

4.6 Proposed Approach

In the following sections, an opinion model, content and people recommender, and opinion estimation strategies are tested in the ABM simulation protocol.

This work aims to fill the research gap with respect to a comprehensive approach to social media simulation concerning content and people recommendation strategies, opinion estimation

and evaluation with respect to a set of metrics that unveil the impact of recommendation strategies with respect to the different strategies adopted.

The adopted model is proposed by Chen *et al.*, [81], which is an adaptation of the DeGroot model [103]. The model also includes backfire [389, 454] and biased assimilation [95]. The former states that individuals are more inclined to accept opinions closer to their own. In contrast, the latter is that, when exposed to the opposite opinions, individuals entrench themselves in their own opinions. As in DeGroot, each node holds an opinion $y(t)$ that is updated in each time step t (that in this case is bounded in the range: $y_i(t) = [-1, +1]$).

The dynamic weights (that determine how influential each neighbour is) are computed as follows:

$$w_{i,j} = \beta_i y_i(t) y_j(t) \quad (4.5)$$

The β parameter is called the **entrenchment** of node i . The parameter captures both the tendency of node i to become more entrenched by opposing opinions and the bias towards assimilating opinions favourable to its own. The opinion of node i in $t + 1$ takes into account the weights and neighbours' opinions, and the formula is given:

$$y_i(t + 1) = \frac{w_{ii} y_i(t) + \sum_{j \in N(i)} w_{ij} y_j(t)}{w_{ii} + \sum_{j \in N(j)} w_{ij}} \quad (4.6)$$

Here $N(i)$ indicates all the nodes in i 's neighbourhood that posted content. The advantages of this model with respect to the original DeGroot model are several:

- Extreme nodes are not stuck in the extremes
- Backfire effect is observed when the disagreement between a node and her neighbours becomes large

A stubbornness parameter is added to the formula to reduce the speed of opinions' convergence. Stubbornness can be intended as the α parameter in the FJ model (see Section 4.4.3), but also the HK model can be adapted to include stubbornness [168, 456]. The higher α , the more stubborn the agent will be, and he will consider the fewer opinions he is exposed to. Equation 4.6 can be adapted in this way:

$$y_i(t + 1) = \alpha y_i(t + 1) + (1 - \alpha) y_i(t) \quad (4.7)$$

Not all nodes change opinion in each timestep. To mimic asynchronous social media, a sample of nodes is selected in each time step, and only these nodes are **activated**. Activation means that nodes post their opinion in their connected nodes' feeds and then read and update their opinion following equation 4.6. The feed also allows the application of different content recommendation strategies, which are explained below.

4.6.1 Network Generation

To obtain a realistic network, the desired properties are:

- **Homophily**: is a property of users' behaviour that implies that new connections (edges) are created preferentially with nodes with similar opinions.

- **Preferential attachments** [194]: usually is referred to the fact that during network morphogenesis, the nodes chose to create an edge *preferably* with nodes with a higher degree.

To obtain these characteristics, the algorithm of Albert-Barabá, [13] is adapted, adding two parameters that govern the preferential attachments:

- Homophily (H): as this parameter increases, it proportionally increases the probability that the node will create an edge with a node with a similar opinion.
- Degree (D): as this parameter increases, it proportionally increases the probability that the node will create an edge with a node with a high degree.

The function to create the network takes as input:

- Number of nodes: is set between 100 and 250
- A tuple with Homophily (H) and Degree (D) parameters, which are fixed for all the nodes.
- A vector that for each node stores his opinion, that are initially randmoly assigned using a Uniform distribution between $[-1; 1]$

The output will be an undirected and attributed graph, the starting point of the ABM simulation.

4.6.2 Recommendation stratgies

This simulation aims to investigate the role of content and people recommender with respect to polarisation dynamics and the diversity of items served. The strategies aim to test if recommending *far or close* content or people affects the outcome of the simulation and to what extent. The concept of *far or close* is operationalised with respect to opinion (content) and the network topology. Content can be considered far if it brings distant opinions. A node in the graph can be considered far because there is a long path between two nodes or because it has a distant opinion.

People recommender

People recommender typically operates by providing a personalised shortlist of users to nodes in the network that each user can accept. If one of the recommendations is accepted, a new edge is created. Sub-strategies are implemented to create variations on the main driver of the recommendation. Each strategy can, in fact, favour or counteract homophily. Following this, opinions are considered while computing the probability that two users will become friends (and create an edge between them). The strategies implemented are the following:

1. **topology based**: leverage the network structure. Given the initial homophily in the network, it is possible to assume that far nodes have, on average, distant opinions.
2. **opinion estimation based**. The first step is the opinion estimation that can be done using the published posts.
3. **random**: a random node j is selected between the ones that do not have an edge with node i .

Strategy (1), (2) have **sub-strategies** that *favour-* or *counteract-* homophily. These two options are at the core of the research because they directly expose the role of recommender systems in creating echo chambers or filter bubbles. For example, the sub-strategy **opinion estimation based counteract homophily** recommends people whose estimated opinion is far from the target nodes.

To estimate opinions and include noise in the simulation, the recommender systems cannot directly access nodes' opinions but must estimate them. Two possible estimation techniques are implemented:

1. Average: the mean of the last 5 posts from a node are used to estimate his opinion
2. Kalman filter: this model allows estimating not only the opinion but also uncertainty about it.

To maintain constant degree distribution and avoid the saturation of the network (a situation where everyone is connected to each other and consequently every shortest path is equal to 1), once a node accepts a new recommendation, one random edge is selected from the pre-existing ones and deleted. This also allows focusing specifically on the role of RS rather than on the opinion model itself.

Content recommender

The nodes' feed is populated with the opinion shared by activated and connected nodes in each time step. Before publication, noise is added to each post. Given all the posts in the feed, the content recommender filters them based on the strategy. The implemented strategies are:

- **Random**: a subset of items is selected randomly from the ones posted by active users.
- **Normal**: Feed is populated with posts generated as normal distribution centred on a specific value for all the nodes.
- **Nudge**: given a specific nudging value (NV), users are nudged toward it, exposing them to a set of content centred on average between the NV and the node's current opinion.
- **Similar**: given a threshold value (TV), items published by friends are selected considering if they fell under the threshold.
- **Unsimilar**: this strategy is the opposite of the similar one. Based on a (un)similarity threshold, only posts with opinions outside it are shown in the feed.

4.6.3 Simulation protocol

In the simulation protocol, the point of interest is to see how the network evolves under different recommendation strategies for information filtering and people recommender. To that end, different parameter configurations were tested. The ABM framework requires defining different parameters that will be tested. Once defined, all the possible parameters combination are executed following Algorithm 1 reported below. Each combination of parameters is executed for T epochs and repeated R times to ensure that results are robust to random parameters initialisation.

Algorithm 1 Execution of the simulation given parameter combination

Input:

- 1: $N = n$ ▷ Number of nodes
- 2: $R = r$ ▷ Runs for each combination
- 3: β ▷ Entrenchment
- 4: $K = k$ ▷ Posts in the feed
- 5: *PeopleStrategy*
- 6: *ContentStrategy*

Output: L (List of metrics for each r, t)

```

7: for  $r \leftarrow 1$  to  $R$  do
8:   while  $t < T$  do
9:     Sample Nodes
10:    Activation
11:    Posting
12:    Content Filtering
13:    Opinon UpDate
14:    People Recommendation
15:    return List of metrics
16:   end while
17: end for

```

Algorithm 1 reflects that at the core of the simulation, there is the goal of better understanding the role of recommendation strategies in favouring polarisation and harmful dynamics. As a baseline is considered the random recommendation of posts and people.

4.7 Results, limitation and future works

The table below reports the metrics used to evaluate the model with relative description. Each metric covers a different aspect of interest: from the emergence of polarization is computed as the sum of the squared difference between each user and mean opinion. The Sarlse bimodality coefficient is computed as $(\text{skewness}^2 + 1) / (\text{kurtosis})$ and indicates how peaked the distribution is. The entropy is used to compute the *diversity* of recommended items (people or content). Disagreement and polarization help us understand how diversified is the informational context of the users. Here is also proposed a metric that evaluates feed satisfaction. This metric is computed taking into account the nodes' parameters that govern the biased assimilation (called β) and the opinion nodes are exposed to in the feed. Firstly the weight of each post in the feed is computed, and the higher the β fewer the weight of the distant posts.

As stated above, the focus is on the different recommendation strategies and to do so, each metric is averaged across the different runs for each recommendation strategy.

Name	Description
Polarisation	Sum of the squared difference between user and mean opinion.
Sarle's Bimodality Coefficient	Based on skewness and kurtosis that takes values between 0 and 1, with the value of 1 corresponding only to the perfect bimodal distribution.
Disagreement	Mean Absolute distance in each neighbourhood
Diversity of Recommendation	Entropy of (binned) opinions in the recommendation list
IntraList similarity	Average cosine similarity of all recommended items.
Feed Diversity	Entropy of (binned) Opinion in the feed
Feed Satisfaction	It values both the feed's diversity and similarity of filtered content.

Table 4.1: The table reports the metrics used to evaluate the model's output

4.7.1 Recommending people

Below are reported metrics to understand the contribution of different **People recommendation strategies**. The simulation for each parameter combination is executed for ($R = 10$) runs, and results are averaged for each metric. Two values of the β parameter, which govern the backfire effect, are presented.

High Beta

With a $\beta = 3.25$, the backfire effect is expected to be strong, given that the higher the β , the higher the backfire effect will be. In the picture below, the dynamic of a random sample of users' opinions is pictured. It is possible to see that for every recommendation strategy, a consensus is not reached, and a strong polarisation characterises the network starting from the early epochs.

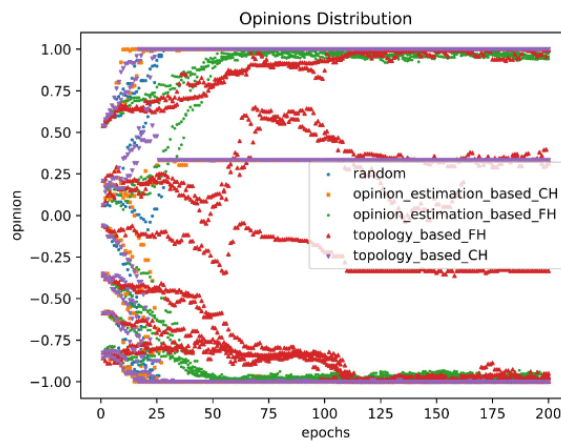


Figure 4.2: Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.

The pictures below report: (i) Feed Satisfaction, (ii) Polarisation, and (iii) Disagreement. The Strategies that favour homophily (Green and Yellow lines) are characterized by a lower **Polarisation** and **Disagreement**. This means that favouring homophily will generate neighbourhoods where opinions become closer (Disagreement decreases) and a slower divergence, given the fact that

polarization in strategies that counteract homophily are always below all the other polarisation curves.

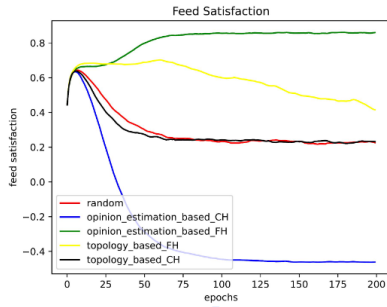


Figure 4.3: Feed Satisfaction is computed as a function that weights close and distant posts based on the increase for all the strategies. β value.

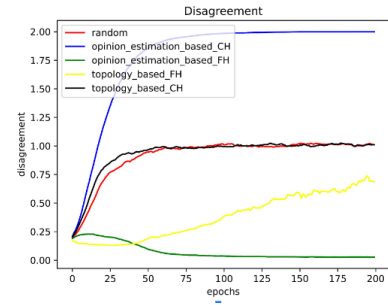
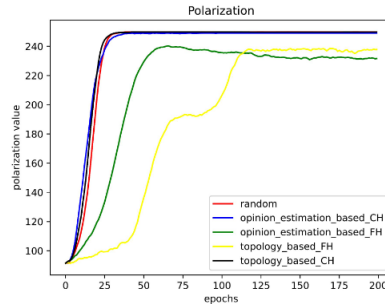


Figure 4.5: Disagreement is a local metric computed in each neighborhood

Low Beta

Lowering the $\beta = 1.25$ allows seeing what happens when the backfire effect is present but with less strength. The figure below depicts the dynamic of opinions for a sample of users. Is it clear that there is a tendency toward consensus for all strategies.

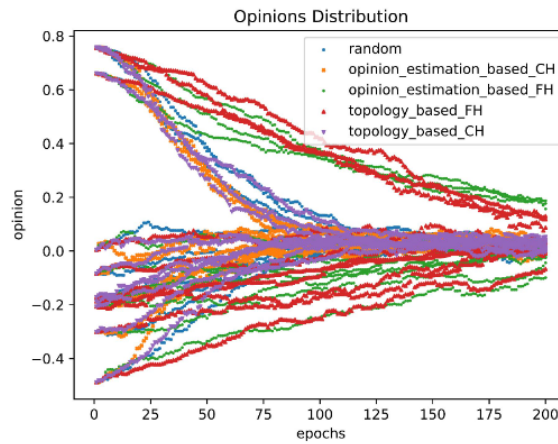


Figure 4.6: Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.

Figures 4.7,4.8,4.9 report the same metrics of previous pictures, but with a lower β , some interesting dynamics emerge. **Feed Satisfaction** is higher for strategies that counteract homophily, while for the opposite strategies, the same metric increase but at a slower pace.

Polarisation too shows a common decreasing tendency, but for strategies that counteract homophily, the dynamic is faster. This seems counterintuitive, but the explanation can be found in the faster convergence of opinion for those strategies, as shown in Figure 4.6. This can be interpreted as the consequence of the fact that when people tend to backfire less, their satisfaction can increase and polarisation be reduced.

Disagreement follows the same dynamic: it increases in the first 25 epochs for the strategies

that counteract homophily and the fastly decreases, reaching levels that are below the disagreement of strategies that favour homophily.

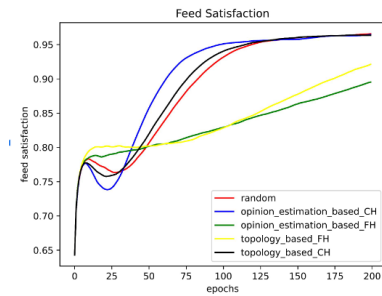


Figure 4.7: Feed Satisfaction.

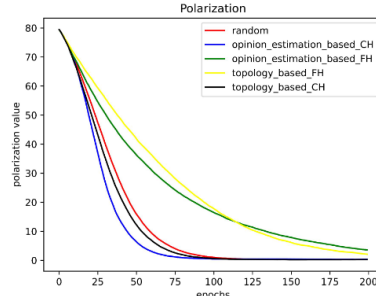


Figure 4.8: Polarisation

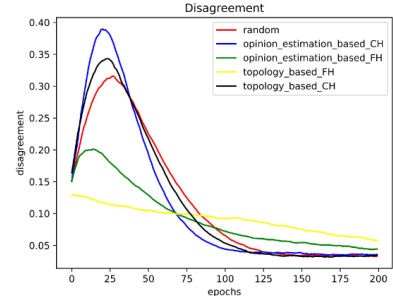


Figure 4.9: Disagreement

4.7.2 Content Filtering

Simulation results with content filtering reveal insights into the polarisation and disagreement dynamics within the given opinion model and network structure. As before, the network presents both homophily and preferential attachments, so nodes tend to be connected to others with similar opinions. Some nodes will become hubs characterised by a relatively higher number of connections. By filtering the content, the goal is to study patterns of agreement and disagreement among participants also toward specific opinions. Additionally, by incorporating satisfaction metrics, it is possible to gauge the level of contentment with the opinions nodes' are exposed to.

The content recommender strategies tested are the ones introduced in Section 4.6.2.

Opinion Convergence

Figure 4.10 below illustrates the dynamics of opinions among a sample of users. It is interesting to see that the **normal** strategy is the only one that appears to have a significant effect on users' opinions, as it is the only one that deviates from the central value. This strategy aims to guide nodes toward the value of -0.4 , and the orange line converges toward this value. On the other hand, the **nudging** strategy fails to reach the desired goal. This can be interpreted as the fact that nudging with a gradual shift is not so effective. In this case, the nudging desired goal of at an extreme (0.9), but nodes converge toward the central value (0.0) Both **similar** and **unsimilar** strategies appear to converge on the central value, suggesting that the content filtering has little effect on the outcome. This suggests that the normal strategy may be the most effective in influencing users' opinions, while the other strategies may be less effective.

Polarisation, Disagreement and Satisfaction

Figures 4.11,4.12,4.13 provide a comprehensive view of the effects of different strategies on key metrics related to opinion dynamics and content filtering. The **feed entropy** plot, computed as the average entropy across nodes' feed, illustrates that the normal strategy has a lower entropy with respect to random but higher than similar, unsimilar and nudge strategies. This can be explained by the fact that posts are similar for all nodes, but it is higher than simial and unsimilar, probably because of the chosen thresholds that limit the diversity in the feed too much. **Feed Satisfaction**

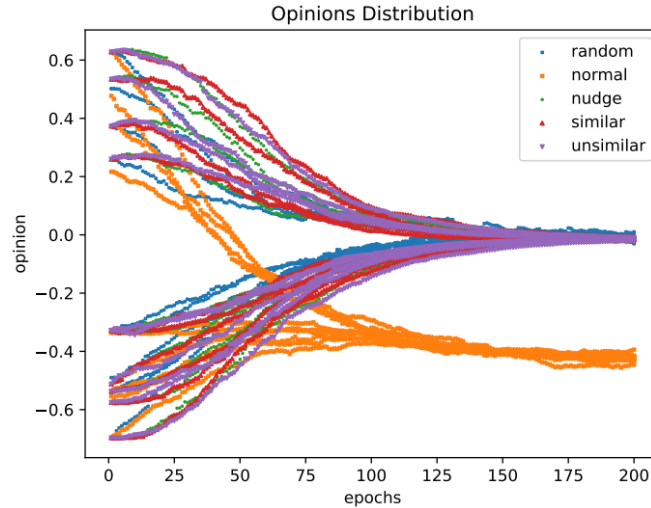


Figure 4.10: Opinion distribution for a sample of nodes. X axis reports the epoch while the Y axis reports the opinion value for each node. Colours represent different recommendation strategies.

among users is the lowest for **Normal** strategy. This can be explained because, for this strategy, all nodes are exposed to similar posts, and the diversity taken into account to evaluate this metric is penalised. The **polarization** plot shows a decreasing trend, indicating that opinions are converging too fast independently from the chosen strategy. The dynamics of polarisation show that opinions converge too fast a viable option to reduce the impact of the opinion model could be to increase the *stubbornness parameter*.

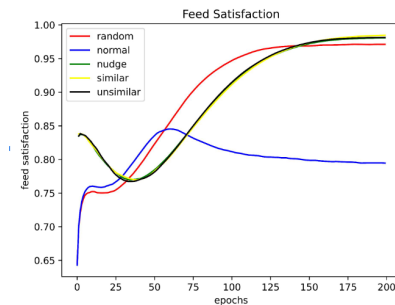


Figure 4.11: Feed Satisfaction represents the utility of the feed.

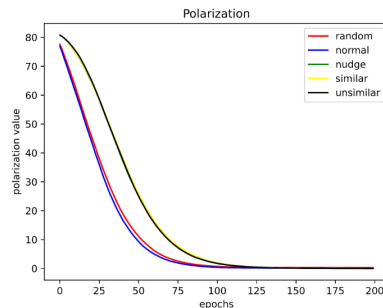


Figure 4.12: Polarisation.

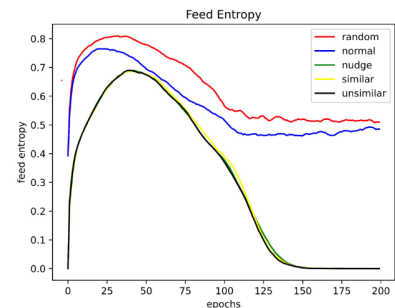


Figure 4.13: Feed Entropy

4.8 Apply real data to ABM simulations

Testing agent-based models of opinion dynamics with real data is crucial for understanding how these models align with real-world phenomena. Validation with datasets gathered from real sources (such as social media) allows for testing model assumptions and parameters and identifying model limitations and areas for improvement. Additionally, comparing model predictions to actual data can provide insight into the mechanisms driving opinion dynamics in a given context. By testing with real data, researchers can increase their models' external validity and enhance their findings' applicability to real-world scenarios. Furthermore, testing with real data allows for explaining the emergence of specific patterns and behaviours observed in the data and predicting future trends.

Datasets are usually snapshots of a temporal graph attributed to nodes and/or edges. Below is a table that lists some real-world datasets that can be used to test agent-based models, along with their characteristics.

Social media	Graph Type	User preferences	People Recommerder	Domain
Twitch [344]	undirected	NO	NO	Streamer
Facebook [340]	undirected	NO	NO	General purpose
Deezer [343]	undirected	Categorical	NO	Music
LastFM [343]	undirected	Categorical	NO	Music
Karate Club [447]	undirected	Categorical	NO	Conflict

Table 4.2: Each row of the table is a dataset, and different features are reported in the columns.

4.9 Conclusion

In conclusion, this chapter has investigated the role of different content and people recommendation strategies in the context of agent-based and opinion models. The model included interesting phenomena such as backfire and biased assimilation with an initial level of homophily and preferential attachments to mimic realistic social networks. The strategies leveraged different aspects of the content and people’s recommendation, from the graph’s topology to opinion similarities. Moreover, recommenders are characterized by imperfect knowledge, meaning RS must estimate nodes’ opinions. Starting from the opinion’s estimation strategy, a step that could be intended as preliminary with respect to serving recommendations can lead to different results in feed satisfaction.

Also different users parameters, such as different levels of β can affect the outcome of multiple metrics for the same strategy. It is also important to note that the role of content filtering is prominent with respect to the one that serves people’s recommendations.

A counterintuitive fact happened with Low beta in opinion convergence. It is possible to note that strategies that favour homophily are the slowest to converge, and this could be caused by the fact that recommending similar users slow down the convergence while showing content and recommending people with a distant opinion force users to converge faster in the case of Lower beta and generate polarization in the case of high beta. This means that the fine-tuning of the β parameter, which determines the entrenchment between users, should be carefully considered also in relation to the specific application because, i.e. specific topics or users could be more prone to backfire and biased assimilation, so it demonstrates a range of flexibility in determining those to phenomena that can help researchers to adapt this reaction based on the specific needs of the simulation.

Results of different parameter combinations and strategies showed that information filtering (both at content and people levels) significantly impacts metrics such as feed satisfaction, polarization and disagreement.

This highlights the importance of careful consideration when designing recommendation systems to minimize negative effects and promote healthy discussion and information sharing.

4.9.1 Limitation and future works

This model has several limitations. First, it only considers unidimensional opinions, while a flow of different topics (which simultaneously coexist) characterizes social media. Another limitation of this study is that the strategies were tested using synthetic data rather than real opinions from different domains. This limits the model's validity, as it is unclear how well it would perform with real-world data. Additionally, this study considered only a few aspects of the content and people's recommendation. It would be valuable to explore further aspects, such as the role of assigning ratings to information items, and introduce propaganda bots in the form of extremely stubborn nodes.

In future work, it would be beneficial to test the strategies with real opinions from different domains to assess the model's validity better and to identify any potential issues that need to be addressed. This will provide a more realistic understanding of the impact of information filtering in real-world scenarios.

Chapter 5

Digital Media Literacy and Information Personalization: A Game-Based Approach for Teenagers

This chapter will address the role of information personalization algorithms in a sequential decision-making setting. Firstly, the potential drawbacks of these algorithms in a decision-making framework will be introduced, and they will be related to the concept known as the *wisdom (or folly) of crowds*. Then, the results of multiple experiments with students are presented. Trials with different classes have been performed inside an educational activity composed of:

1. Digital media literacy educational talk
2. A Game-based experience that mimics information personalisation during a repeated estimation task

5.1 Information personalisation and decision making

Information personalisation is a valuable tool in an information environment characterised by an abundance of news, videos, people, pictures and songs. At the same time, the outbreak of the COVID-19 virus and the global action taken by the World Health Organisation (WTO) highlight the fact that another emergency calls for action: *infodemic* [449].

The term coined by the WHO [299] focuses on the relevance that a set of *good practices* of scientific communication to a broader audience must be taken into account to avoid the spreading of non-scientific therapy and fake news. COVID-19 is the perfect example because information affects actions in that case, especially for vaccine adoption. After all, West countries' governments only partially opted for mandatory vaccination and preferred adopting a persuasive education campaign [379].

Empirically, [87] crawled and analyzed tons of data from different social media platforms and found that the spreading patterns and engagement levels in part depend on the specific platform and the specific population with its own preferences distribution.

As mentioned in 3.1.5, users' opinions in a decision-making setting can, given a mechanism of incentives and punishments, skew and modify users' decisions, as in the case of *Information Ger-*

rymandering [377]. The authors demonstrate that the informational context affects users' actions and decisions with respect to estimation tasks, leading to undemocratic decisions.

5.1.1 Digital Media literacy and remote teaching impact

The advent of digital platforms leads researchers to propose new forms of literacy, including these digital technologies. Given their pervasiveness in people's daily life and the relevance of a correct evaluation of content, users are exposed online, especially teenagers that must develop competencies to use these tools.

The possibility of learning also outside schools at a low cost allows social media and digital tools to overcome the physical limitations of in-person teaching, especially during COVID-19 outbreaks when state authorities announced the end of face-to-face teaching in primary, secondary and higher education. This forces institutions to adapt their offer to new paradigms of online learning.

The need for a digital platform that can be a substitute for face-to-face teaching in an unpredictable emergency such as the spread of a virus is needed [292] to ensure at least in part a continuum of students' learning path.

Teachers can also benefit from using social media such as Twitter, where they can find resources and opportunities to share opinions and experiences with no cost and little effort [277].

The role of remote teaching has been addressed systematically only after the COVID outbreak. Results vary based on the age of students, but it is clear that there are multiple drawbacks to an only-remote-teaching system. Surveying around 400 students from the physics department in Italy [252] found that even if students appreciate the organization of remote teaching activities, this modality affects engagement and results.

The introduction of social media to create a better learning environment can be fostered by using it as a motivational and engagement-boost tool. Main results from Escamilla *et al.*, [121], indicate that TikTok, at least for university students, could be helpful.

5.2 Wisdom of folly of the crowd?

In the finance sector, one of the most expensive pieces of information is what professional investors expect in the future. People used to pay to see what a forecaster thought about the next market's fluctuations. A cheaper solution could be to average the opinion of a group of everyday people and see what happens. It could turn out that the average opinion of ordinary people can outperform individual experts [333]. This means that the knowledge, which any person has at the individual level, can emerge from the crowd, leading to the so-called *wisdom of the crowd*. Two hypotheses can explain this phenomenon. First, if the opinion sample is hypothesized to be *Independent and identically distributed (i.i.d.)*, it is possible to expect that errors will compensate for each other. Second, if the sample size is increased asymptotically, the correct answer will be reached. Given the fact that almost everything is marketable, an increasing number of platforms based on *Web3.0* offers the possibility to trade the outcome of events ¹.

¹More info here: <https://polymarket.com/> or <https://www.yolorekt.finance/>

5.3 Related works

5.3.1 Digital Media Literacy and Teenagers

Teenagers represent the most common social media users. Education systems are starting to introduce digital media literacy in their curriculum [70, 357] to educate students about digital citizenship as the proper and responsible way of using digital technologies [335], enabling users to critically approach social media and deal with its threats [150, 258, 277, 320, 393].

Understanding the impact of social media on their mental health and well-being is crucial [210, 296] also because multiple studies highlight the strong and negative correlation between the usage of social media and self-reported mental health, suicidal intention, etc. Tsitsika *et al.* [405] also found a positive association between heavier social media use (more than 2h/day) and anxiety and depression for a sample of European students. A systematic review confirms a general correlation between social media use and mental health problems but highlights the complexity of this relationship [204].

During the COVID-19 pandemic, social media allowed locked-down individuals to maintain connections; consequently, the average time spent on these platforms increased. But as shown by Parlak *et al.*, [309], the increase in time spent on social media by teenagers is a good predictor of their COVID anxiety score, finding also a correlation with social media addiction.

Moreover, specific social media features, such as unfriending, can accelerate the emergence of the so-called echo chambers [353]. As shown by Cinelli *et al.*, [86], Facebook is found to be the social media platform where homophily and information diffusion generate the most correlated neighbourhoods in terms of opinion leaning. Bail *et al.* [30] find that exposing Twitter users to opposite political views increases polarization.

A digital literacy activity was proposed by Choolarb *et al.* [84] that merges both an augmented reality tool and a gamified strategy. Results show that, in terms of satisfaction and learning goals, the experimental group (pupils who get the augmented reality and game-based learning program) increased their performances with statistically significant differences in their evaluation with respect to the control group. A systematic review [90] finds that cyber-security literacy is mainly framed as a web-based game by government agencies, non-profit companies, and academic institutions. These activities are mainly composed of a gamified quiz or worksheet. Increasing the interactivity level of media literacy initiatives is beneficial to improve a correct reliance on machine learning models [83].

5.3.2 Recommender Systems and Social Media

As highlighted in chapter 2, platforms leverage machine learning systems trained using billions of examples taken from users' logs or other data sources to efficiently provide social media users with the most engaging content and needed information. These algorithms classify, filter, and recommend content or other people in a personalised manner to each user. This personalisation step can belong to the Collaborative Filtering family, where users with similar interests are leveraged to find potentially interesting items for each user [354]. Alternatively, recommender systems can leverage content representations (in the form of numerical embedding or any feature that can be used to compute a similarity between two items) to find items similar to the ones liked or purchased by the user. Still, while maximising users' engagement, social media platforms may harm other stakeholders that use the platform [66, 263]. Their recommender systems may exasperate users'

tendency to connect and populate their virtual environment, or social sphere, with like-minded users. They may also be served with content they already like [86, 298]. Moreover, if users are unaware of the consequences of information personalisation, their perception of reality may be more strongly distorted [408].

Boeker and Urman [54] pave the way to a better understanding of TikTok’s recommender system. Findings show that certain factors influence the recommender system more than others and point out how explicit factors (e.g. language, location, previously liked pages and creators) influence the recommender more than implicit factors such as the video view rate. It is crucial to highlight that ByteDance, TikTok’s owner, had never released or opened to a third-party audit their recommender system to establish which, for example, are the relative weights of explicit and implicit feedback taken into account by their recommender, that is incidentally one of the main reasons of TikTok’s success [452].

Besides the influence of algorithms, users can also reciprocally influence each other and the recommender systems can accidentally amplify negative dynamics. For example, they may increase or decrease the visibility of emotionally loaded content, but as shown by Kramer *et al* [219] that studied the relation between users’ exposure and user content production, when the exposition to positive expressions is reduced, people produce fewer positive posts and more negative posts. Furthermore, emotional and sentiment load in messages can itself affect the spread of content, as shown by Brady *et al* [59] through modelling the diffusion of moral-emotional language in political content. These results indicate that emotions expressed by others on Facebook influence our own emotions and highlight that emotion contagion can take place on Facebook. Moreover, even if the exact role played by recommender systems and user selection in the creation of echo chambers isn’t clear [31], it is crucial to enrich media literacy with specific notions and strategies to deal with these systems and issues [408].

5.3.3 Wisdom of Crowds

The phenomenon of the Wisdom of Crowds was discovered by Galton in 1907 [138], and a century later social media are challenging his findings. Forecasting and estimation tasks can get better accuracy from crowd-sourced answers rather than a single expert [130], and a crowd can be smarter using a collective decision making deliberation system [241, 283].

The reasons why this can happen are grounded on the hypothesis of uncorrelated errors and enough diversity in the initial distribution of answers, so that even if the individual answer is far from the right one, an aggregation of answers will be more accurate [301]. It is clear that the Wisdom of Crowds is closer to a statistical phenomenon rather than a social feature, but social influence can undermine the hypothesis upon which this statistical phenomenon relies on.

This collective decision-making process, where participants are aware of others’ estimates or answers and they can revise their answer can fail for many reasons. Social influence can reduce the efficacy of the wisdom of crowds given the fact that multiple factors can lead to correlated errors.

To address this interplay between social influence and individual conviction, Mavrodiev and Schweitzer [254] derive analytically the condition with respect to the initial error and diversity of opinions that determines an increase or decrease in accuracy under a decision model that includes social influence and individual conviction.

The effects of network structure on the performance of wisdom of crowds were investigated by

Becker *et al* [42] that proposed an experimental setup where two radically different but simple graphs are created. The first type of network was a centralized one: all the nodes were connected with only a few other nodes that had disproportionately larger number of connections, generating a sort of a star graph, so all the nodes except the central ones have a degree equal to one. The decentralized condition was defined as networks where everyone has the same number of edges, i.e. connections. Participants were involved in a repeated estimation task, and the results confirmed the existence of the wisdom of crowds, because for the first estimate the mean answer was more accurate than that of the majority of the participants' individual answers. Also, between different estimates the mean of the estimates was more accurate and there was a reduction of the standard deviation. Furthermore, findings show a huge impact of network structure: the decentralized network has a positive effect on the overall accuracy while the centralized one is always skewed toward the direction of the central nodes. Analytical approaches such as that of Mavrodiev and Schweitzer [254] also argue that the wisdom of crowds is really a weak phenomenon and a small dose of social influence can heavily reduce the diversity of answers without ensuring an increase in accuracy. Moreover, the increase in confidence, i.e. the reduction in standard deviation, may be caused by a convergence that is not related to an increase in accuracy [254].

Literature shows that social influence and network structure deeply affect the wisdom of crowds, but individuals may still have other possibilities to coordinate and improve. Navajas *et al* [283] study the role of group deliberation in parallel with social influence. The authors create a repeated estimation task that involves many individuals and then let them form small groups just after individual estimation to allow the group to reach a consensus and deliberate. Results show that if individuals are allowed to form small groups after the first estimation the possibility of deliberation within groups improved the crowd's collective accuracy.

The role that political bias and partisanship could play with respect to the wisdom of crowds has been investigated too. Individual motivations (e.g. political preferences) can affect the reasoning [223] and shape opinion about facts, so it is crucial to also test the wisdom of the crowds hypothesis outside the context of nonpartisan estimation tasks commonly used for experiments.

Becker *et al* [43] create political homogeneous networks and then perform a repeated estimation task to test whether the social influence and political partisanship can affect the accuracy of non-partisan estimation tasks and if during the estimation beliefs became more extreme instead of more accurate. Findings show that homogeneous networks are resilient to the propagation of partisan bias, so the wisdom of crowds in networks is robust to political partisan bias. The effectiveness of the wisdom of crowds has been tested also with respect to fact-checking by Allen *et al* [17]. In this work, more than one thousand participants have been recruited using Amazon Mechanical Turk, and findings show that a small, politically balanced crowd of people can match the fact-checkers' accuracy and agreement in the task of labelling news articles.

5.4 Experimental Methodology

The proposed experiment entails a game-oriented social estimation task inspired by the wisdom of crowds [138, 241] inside an educational activity aimed at raising awareness about social media influence and information personalisation effects. The experiment is preceded and followed by questionnaires. In the very initial step, a random and unique code was assigned to each participant

to ensure data anonymity. In the pre-treatment questionnaire, participants are asked for socio-demographic questions and social media usage information. Both in the pre and post-treatment surveys participants are asked, using two items (namely: **Perceived Influence on themselves** and **Perceived Influence on peers**) on a 6-point Likert scale which is the perceived influence of social media. The activity was designed to last no longer than an hour and a half to avoid the participants becoming mentally fatigued when filling out the final questionnaire. Each step of the protocol is depicted in Figure 5.1.

First, a digital media literacy talk was given to students, after which participants were involved in a game organized as a repeated estimation task. In the game, different information scenarios were tested [42, 283]. In the first, between trials, the students were given the correct aggregate of the crowd's estimate, while in the second they were presented with a biased one, i.e. skewed towards the wrong answer.

The intuition is that through direct exposition to one of the most impacting echo-chamber and filter bubbles consequences, i.e. when biased sampling distorts users' unbiased opinions, followed by its explanation, the students can reason about the role that information personalisation has had in misleading them and will become more aware of these mechanisms and their effects. To discuss the results the metaphor of a "recommender system that selects the responses most relevant to the majority of users" adopted in the game is also stressed.

For comparison, a baseline educational activity that comprised only the digital literacy talk was also performed. Given the pandemic situation, the educational activities were performed remotely, with students present in the classroom, and researchers connected through video conferencing software.

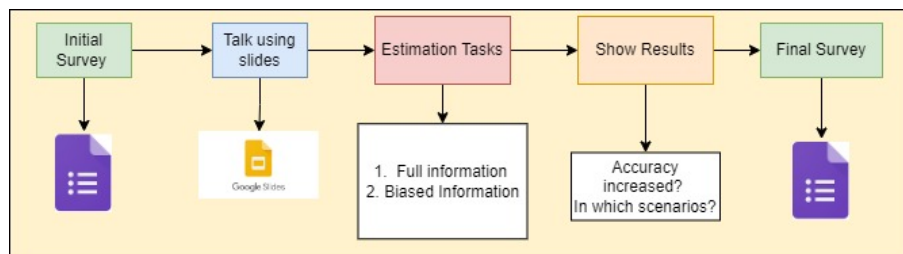


Figure 5.1: The proposed experimental protocol. The total expected time for completion of the activity is 1h and 30 minutes.

5.4.1 Participants

The activity targets specifically teenagers enrolled in high schools. A pilot experiment was performed for the first time involving Computer Science master's students from the University of Milan-Bicocca. The sessions took place in the high schools ITET G. Caruso in Alcamo (Italy) and in Liceo Morgagni in Rome (Italy). All participants were between 18 and 21 years old ($N=52$, Mean = 18.7, Standard deviation = 0.85).

5.4.2 Digital Media Literacy Talk

The digital media literacy activity has been developed to raise students' awareness concerning social media threats and empower them [393]. The goal of the digital media literacy talk is to

contextualize the relevance and impact of information personalisation [408] and its applications such as recommender systems to point out their potential pitfalls and biases [77].

The digital media literacy talk covers the differences between traditional media and social media with their complexity and pervasiveness, the impact of cognitive biases, and, finally, their interplay with information personalisation algorithms, highlighting the concepts of echo chambers and filter bubbles. During the talk, slides were used as a teaching tool.

The structure of the talk is detailed below and in Figure 5.2 a sample of the slides shown to students is reported:

- **Disintermediation:** a pivotal player in social media, the prosumer (both a consumer and producer of content) is introduced to the user. New media do not have an editorial board and content can freely spread around the network.
- **The Network Structure:** different graph structures to highlight the different *roles* that nodes can have as bridges for information spreading between friends are introduced and how each user affects their friends through their behaviour (liking, sharing). It is also shown that the network dynamics are complex to predict, as demonstrated by the unexpected popularity of some content.
- To introduce **virality** [326] the narratives and stereotypes that are more common in viral content are introduced, but was also highlighted how this model can negatively affect users, especially teenagers who aim to get popular on social media.
- Information personalisation is introduced as a helpful tool to overcome information overloading. An intuitive description of the principles of recommender systems is given to students to better understand the concept of filter bubbles.
- Echo chambers are explained as a difficult to observe phenomenon that may lead to social fragmentation and polarization and may be exacerbated on social media given the partial and skewed representations provided to users by information personalisation systems.
- The notion of backfire is introduced as a way algorithms can increase polarisation when exposing users to opposite and extreme content causing a vigorous reaction in users.
- Cognitive bounds [326] and biases, e.g. homophily were also present that may be causes alternative to algorithms of polarization and other issues observed on social media.
- Conclusive remarks regarding the opacity and the lack of transparency by social media companies were presented.

5.4.3 Wisdom of crowds game experience

While the **baseline scenario** ends after the digital media literacy talk, in the **experimental scenario** the wisdom of crowds inspired game followed, and a brief introduction was given to the participants. The game is executed as a repeated visual estimation task with an information personalisation step. The goal for each participant is to give the right answer with the possibility to revise their answer once after being presented with social information, e.g. in the form of an

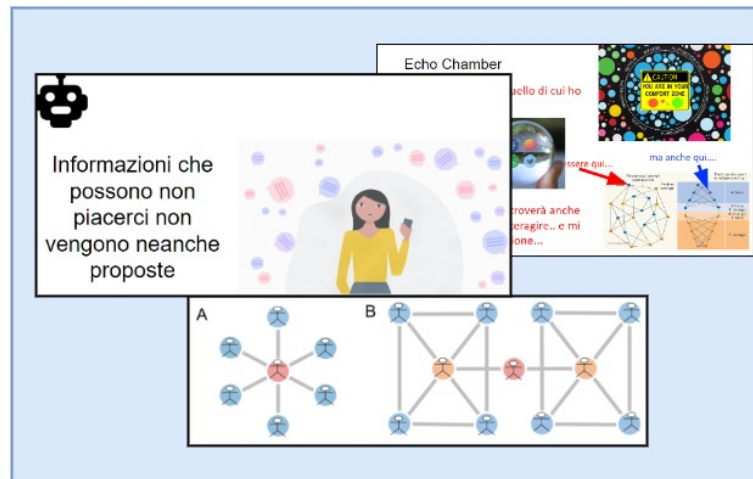


Figure 5.2: Sample of the slides presented to participants.

aggregated biased or unbiased peers' votes [283]. The hypothesis underlying information scenarios is that the unbiased scenario is the one where *biased recommendations* are provided and it is here that participants experience the pitfalls of information personalisation.

The first step of the WOC experience is to show students an image populated with red dots in random locations over a white background, as depicted in the top right of Figure 5.3. This task is independent of the level of individual knowledge, different from the study by Navajas *et al* [283].

The second step of the game is the **individual estimate**. Here students can choose their answer between different ranges of values. The social information displayed to them is an aggregation of all the answers for that trial, and a histogram of all participants' answers is chosen to be shown.

In the **correct feedback** condition, the aggregate is taken of the whole set of answers, while in the **biased feedback** condition, the aggregate is taken over a subset that exasperates the current error. The metaphor here is that this biased set is selected by a black-box recommender system that selects the responses most relevant to the majority of users.

After the information disclosure participants can revise their estimates and they have the possibility to give a second individual answer. A total of four trials with each batch of classes, where half of them were correct and the other half biased were performed.

In terms of users' response precision, according to previous studies [42, 241] it is expected that in the correct feedback condition the second estimate will be more accurate than the first [283], while in the biased feedback condition, it is expected to decrease. In Figure 5.3 the protocol of the game is reported.

5.5 Results and discussion

The data reported referring to multiple sessions, one for baseline and another 5 for the full experiment, that involved separate classes of students.

In the picture below the 6-point Likert scale item (influence on themselves) is reported) for all the experimental-group classes is reported. It is clear that there is shift in the distribution.

To understand if there are differences a *the t-test on two RELATED samples* is computed. The t-test check for the statistically significant difference in the mean of the two samples. In particular,

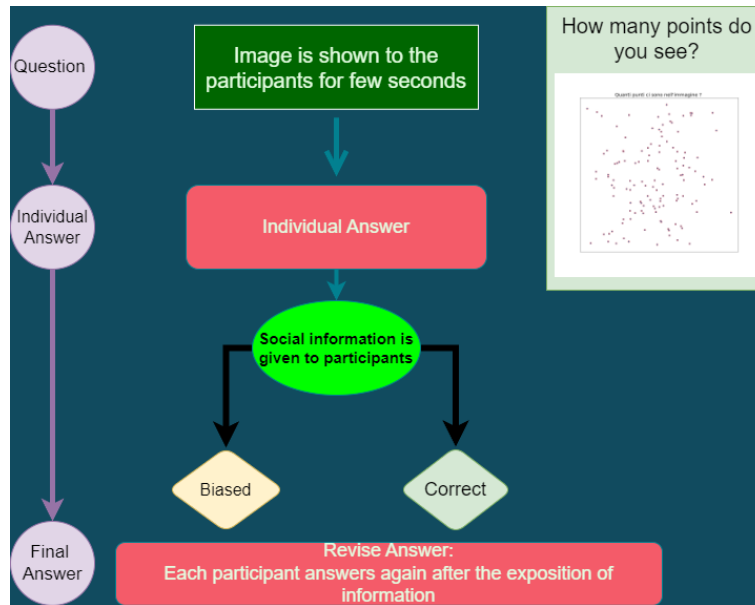


Figure 5.3: Structure of the game-based wisdom of crowds estimation task. It starts with the question where the image on the top right is shown and they are asked to estimate the number of dots in it. After the aggregate of individual answers, as social information, is disclosed to the participants they can provide the final answer.

Quiz: How much do you think social media can influence your opinion?

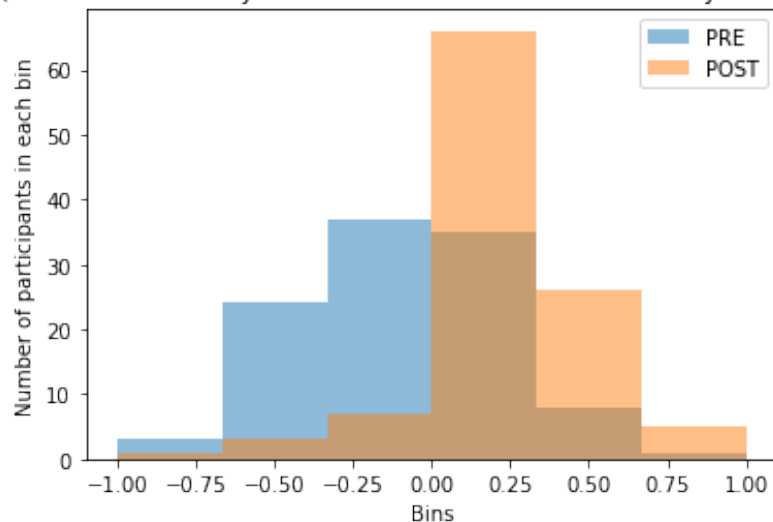


Figure 5.4: The blue histogram reports the **PRE-activity** questionnaire, while the orange one the **POST-activity**.

	How much do you think people are influenced by social media?	How much do you think social media can influence your opinion?
p-value	0,5	1,00E-13***

Table 5.1: Perceived Social Media Influence before and after the intervention. Columns represent the two items: (I) Influence on them and (ii) Influence on their peers. **P-Values** are reported and * means the significance level at 0.005

the alternative hypothesis states that the mean of the distribution underlying the first sample (**Pre-survey**) is less than the mean of the distribution underlying the second sample (**Post-survey**). The table below reports the p-value for both items, namely influence on themselves and on their peers.

In Table 5.2 the average answer for both scenarios and both items is reported in the initial and final questionnaires. Only in the **Experimental** scenario, when results are shown to students, there is an increase in the average perception.

Perceived Social Media Influence (Number of Participants)				
Target	Self		Other	
Phase	Initial	Final	Initial	Final
Experiment	2.89 (19)	3 (18)	3.74 (19)	3.89 (18)
Baseline	3.08(32)	2.66 (32)	3.86 (32)	3.69 (32)

Table 5.2: Perceived Social Media Influence before and after the intervention. Columns Self and Other represent the target of the question related to the perceived influence of social media on themselves or their peers. Values in cells correspond to the average, while the sample size is reported in brackets.

For each session (Experimental and Baseline) and each item proposed (perceived influence of social media on them or their peers) the difference between the initial and final answer for each participant is computed. Then the number of participants with an increase (positive difference) and decrease (negative difference) between the initial and final item is computed. Then for each item about perceived social media influence (on themselves or their peers) Fisher's exact test is computed and p values are reported in Table 5.4.

In Table 5.3 the contingency table for each item proposed to students, namely the perceived social media influence on themselves and on their peers, is reported and used as input for Fisher's test.

It is clear that in the experimental condition, a great majority of participants showed a positive difference between the final and initial survey in both items (perceived social media influence on themselves and perceived social media influence on peers).

The Fisher test for each perceived influence item considering the not complete answers is computed too (left column of Table 5.4). The p-values allow rejecting the null-hypothesis that the two classifications are not different with a significance level of 0.05.

	Influence on self		Influence on peers	
	baseline	experiment	baseline	experiment
increase	10	9	11	10
decrease	16	3	14	2
not answered	3	13	3	13

Table 5.3: The table reports the contingency tables for each item proposed (Perceived Influence on themselves and Perceived Influence on peers). In each row, the number of participants with an increasing (decreasing or not answered) difference between the initial and final survey is reported for each condition (baseline and experimental).

In the baseline condition, even if the mechanism and drawbacks of social media are explained to the participants, the traditional educational activity alone is not capable of increasing the perceived influence

<i>P values</i> Significance Level .05		
	Both Answers	Not complete
self	0.000596	0.015742
peers	0.000441	0.00604

Table 5.4: The table reports the p-values for each item in the survey (rows). Columns report both cases where unanswered questions are either considered a decrease or a class.

of social media. The proposed game is hypothesized to be a good metaphor for the black box recommender system's mechanisms. In the proposed setting, users have no control over it and are not informed about the adopted strategies.

5.6 Conclusion

In this paper, a Digital media literacy activity composed of an educational talk alongside a game-oriented strategy to increase the efficacy of educational interventions regarding social media threats [88] has been described. The aim is to boost awareness of social media threats allowing students to directly experience phenomena such as echo chambers, and filter bubbles and their consequences that may be exacerbated by automated systems such as recommender systems.

Results presented in Section 5.5 showed that a game-based direct experience, inspired by the wisdom of crowds phenomenon [42, 283], can increase the perception of social media influence on participants with statistically significant results if performances of the game are shown to students. The Wisdom of the crowd effect is expected if the correct aggregated estimate is shown to participants in line with the findings by Becker *et al* [42], but due to social media influence, a biased aggregation of crowd estimate reduces the overall accuracy [241]. The difference between pre and post-surveys demonstrates that, at least in the short term, experiencing the echo chamber directly increases the perception of the influence of social media on their opinions and decisions.

5.7 Limitation and future works

Different statistical tests were conducted to check the effectiveness of the activity with promising results but also, several limitations. Firstly, the sample size is too small to generalize findings, and in particular future studies should investigate if results are robust across different countries. Secondly, the effectiveness of the activity in the medium and long term is still unclear. Lastly, participants' behaviour during the game and the magnitude of social media influence could be better investigated in relation to other scales such as the Fear of Missing Out (FOMO) [72, 324] that has been linked to problematic social media use and negative health outcomes among adolescents. In future works, the integration of the activity as a web tool for educators could be a solution to promote its diffusion.

Chapter 6

A Deep Learning Approach to Identifying Harmful Content using Graph and Word Embeddings

This chapter presents a deep learning model that combines a graph-based neural network with a pretrained transformer language model, specifically Graph Attention Convolution and ELECTRA, respectively. The proposed model was designed to identify harmful tweets related to COVID-19, specifically focusing on subtask-1C of the CheckThat!Lab shared task, which was part of CLEF 2022 ¹. In this binary classification task, the model achieved a binary F1 score of 0.28 on the test set, outperforming the official baseline by 8%. One of the key strengths of the proposed approach is its ability to incorporate various forms of information, such as social context, through the use of graph layers. In the case of identifying harmful tweets related to COVID-19, graph networks can help to incorporate social context by capturing the relationships between users, such as followers, retweets, and mentions. This allows the model to learn from the collective behaviour of users and their interactions, rather than just individual tweets. In addition, the proposed model employs a pretrained transformer language model, ELECTRA, which has been trained on a massive amount of text data and can effectively capture the semantic meaning of texts. By combining graph machine learning with a language model, the proposed approach can effectively capture both the social context and semantic meaning of tweets. This architecture suggests that the proposed approach can effectively identify harmful tweets related to COVID-19, which could be useful in preventing the spread of misinformation and promoting public health but also flexible enough to incorporate different types of information.

6.1 Introduction

Throughout the COVID-19 outbreak, the spread of misleading information online related to news on the pandemic could be observed, for example, on social media. The three tasks proposed for the CheckThat!Lab@CLEF2022 [279, 280] aim at addressing related issues, namely:

1. Identifying relevant claims in tweets
2. Detecting previously fact-checked claims
3. Fake news detection

This chapter aims to propose a model to address the first task [278]. Furthermore, this task includes four different subtasks. Subtask 1A is about determining check worthiness of tweets (i.e. given a tweet,

¹More info here: <https://clef2022.clef-initiative.eu/>, retrieved on April 7, 2023

predict whether it is worth fact-checking). Subtask 1B is about detecting verifiable factual claims: given a tweet, predict whether it contains a valid factual claim. In Subtask 1C, the focus is on harmful tweet detection: given a tweet, predict whether it is harmful to society and why. Finally, Subtask 1D is related to attention-worthy tweet detection: given a tweet, predict whether it should get the attention of policymakers and why. This task is defined with eight class labels. For subtasks 1A and 1C the official evaluation metric is the **binary F1 score** with respect to the positive class; for subtask 1B the metric used is the **accuracy**, and for subtask 1D it is the **weighted F1 score**.

The proposed model is for **Subtask 1C - English language**. The model takes advantage of an ELECTRA-based document embedding and a text graph that is processed using a Graph Convolutional Network (GCN). The goal is to introduce a novel method that can handle different types of heterogeneous textual or social information. It is shown how a first version of such a model performs on the proposed task, leaving room for improvements in future research in the domain. To support reproducibility and future research directions, the Code is publicly available ².

This chapter is organized as follows: Section 6.2 presents some related works about using deep learning methods for similar text classification tasks. In Section 6.3, the approach is described in detail, explaining each model layer’s choices and configuration. In Section 6.5, the results obtained on the official test set are reported. In Section 6.6 some interesting future directions are investigated before closing the chapter with concluding remarks in Section 6.7.

6.2 Related work

Motivated by the notable performances reached in various text classification tasks, where deep AI models [217, 364, 366] outperformed classic techniques used in natural language processing (e.g. Bayes, Decision Tree, K-Nearest Neighbour, Support Vector Machine) as also reported in [171, 435], a deep learning-based approach have been chosen.

The proposed model is based on a transformer-based document embedding and a GAT (i.e. Graph Attention Layer) applied to a text graph representing the structure of each tweet. The transformer-based embedding **ELECTRA** is adopted, a language model presented by [89]. ELECTRA does not mask the input as BERT [107] but replaces some tokens with plausible alternatives sampled from a small generator network. Then, instead of training a model that predicts the original identities of the corrupted tokens, a discriminative model is trained to predict whether a generator sample replaces each token in the corrupted input or not.

GCNs (Graph Convolution network is a term used as an umbrella-term to include the different graph-based neural network layers) take advantage of graph data structures made of vertices and edges (also called nodes and links). Nowadays, graphs are used in many real-world applications, including traffic prediction [231], computer vision [321], social networks [28, 365] and many more.

GCNs for text classification are discussed in [443]: the authors propose a novel graph neural network method and model a whole corpus as a heterogeneous graph to learn words and document embeddings jointly with graph neural networks. Results on several benchmark datasets demonstrate that the proposed method outperforms state-of-the-art text classification methods without using pre-trained word embeddings or external knowledge. The model proposed also learns predictive word and document embeddings automatically.

In [233], a more sophisticated approach is proposed. In particular, the authors discussed a flexible Heterogeneous Information Network (HIN) for modelling short texts. The model can integrate additional information and capture their relations to address semantic sparsity. The authors propose heterogeneous graph attention networks to embed the HIN for short text classification based on a dual-level attention mechanism, including node-level and type-level attention. The attention mechanism can learn the importance of different neighbouring nodes and the importance of different node (information) types to a current node.

²<https://github.com/sagacemete/CLEF2022CheckThat.git>

A broader overview of a range of text classification applications using GCNs is given in [455].

6.3 Proposed model

The model architecture with input and output shapes of each layer is shown in Figure 6.1 along with parameter distributions of each layer. The proposed model is composed of two modules:

- Graph creation and embedding
- Pretrained document embedding

```

-----
Model Parameters
-----
Layer.Parameter      Param Tensor Shape      Param #
-----
conv1.att             [1, 4, 450]             1800
conv1.bias            [1800]                  1800
conv1.lin_l.weight    [1800, 815]             1467000
conv1.lin_l.bias      [1800]                  1800
conv1.lin_r.weight    [1800, 815]             1467000
conv1.lin_r.bias      [1800]                  1800
conv2.att             [1, 1, 450]             450
conv2.bias            [450]                   450
conv2.lin_l.weight    [450, 1800]             810000
conv2.lin_l.bias      [450]                   450
conv2.lin_r.weight    [450, 1800]             810000
conv2.lin_r.bias      [450]                   450
lin1.weight           [609, 1218]             741762
lin1.bias             [609]                   609
lin3.weight           [2, 609]                1218
lin3.bias             [2]                     2
-----
Total params: 5306591
Trainable params: 5306591
Non-trainable params: 0
-----
Model Architecture
-----
GCN(
  (conv1): GATv2Conv(815, 450, heads=4)
  (conv2): GATv2Conv(1800, 450, heads=1)
  (lin1): Linear(in_features=1218, out_features=609, bias=True)
  (lin3): Linear(in_features=609, out_features=2, bias=True)
)

```

Figure 6.1: Model parameters (Top) numbers in brackets indicate parameters’ tensor dimensions; the last column indicates the number of parameters in each layer. Model architecture (bottom) model input and output shapes in each layer (figure taken from the Google Colab notebook).

Geometric deep learning [61, 214] has led to a growing number of new architectures as well as novel applications, including text modelling [116]. The representation of each tweet as a graph starts with text preprocessing (the set of techniques used to clean and prepare text data before it can be used for analysis or machine learning. This may involve tasks such as tokenization, stop word removal, stemming or lemmatization, normalization of text, handling of special characters or formatting issues), and Part Of Speech (POS) tagging, a natural language processing (NLP) task that involves assigning a grammatical category, such as noun, verb, adjective, or adverb, to each word in a sentence. After these steps, each unique tagged word in the tweet corresponds to a node in the graph, and the adjacency matrix is populated connecting each node with all words in a window equal to 3. Each node is annotated with various features discussed in 6.3.2. The proposed architecture is composed of two graph attention convolution (i.e. **GATV2Conv**) layers proposed by [60]; the node-wise representation outputted by the GATV2Conv layers is passed to a max pooling operator and a dropout layer. The output is concatenated with the document embedding generated using ELECTRA [89]. Finally, two dense layers, and a rectified linear unit (ReLU) activation function between them, output the model predictions for each class.

Before discussing the network architecture and hyperparameter settings, it is worth mentioning that each dataset split (training and test per language) consists of individual tweets and their corresponding labels.

6.3.1 Graph creation

The graph module takes as input a raw sample (tweet) and outputs the tweet represented as an undirected, attributed graph. In Figure 6.2 each step of the preprocessing pipeline is depicted. The custom preprocessing function uses the python NLTK package [50]. Below all the preprocessing steps involved are listed:

- *Lowercasing.* This step is used to get the same embedding, e.g. for the words **Hello** and **hello**
- *Removing Stopwords.* Generally speaking, stopwords are used with high frequency, but in many cases, they are not really informative, e.g. prepositions and articles belong to this category.
- *POS Tagging.* In this step, each word in the tweet is classified into its parts of speech class and labelled accordingly using a one-hot encoding. These vectors correspond to the respective POS tag out of all 43 POS classes in the NLTK package.
- *URL removing.* All URLs in each tweet have been removed.
- *Hashtag symbol and tagged accounts.* All hashtag symbols have been removed along with tagged users.

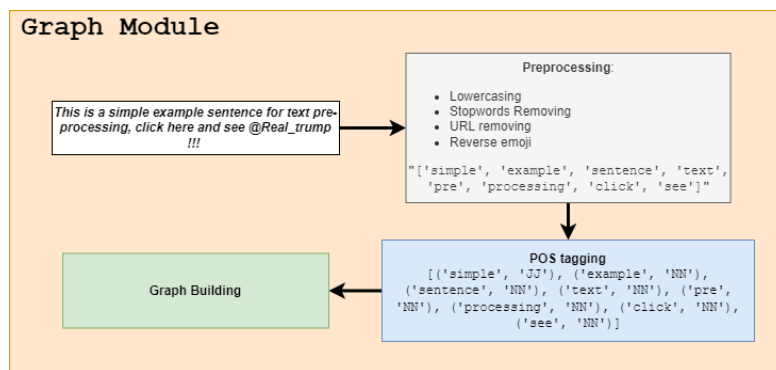


Figure 6.2: Graph representation: each tweet is represented as a graph after pre-processing and POS tagging

Starting from the output of POS-tagging, a strategy that associates an edge to each word with all words in a window equal to 3 is adopted. If a word is repeated more than once, only the first occurrence is considered a node, while edges are updated accordingly. Edges are unweighted.

6.3.2 Node characterisation

As mentioned above in Figure 6.1 each node is characterized as an 815-dimensional vector. The first 768 features correspond to the pretrained ELECTRA document embedding obtained using the FLAIR package and applying the introduced preprocessing steps ³[9]. Different transformer-based models using the tweet text data as input have been evaluated to select the best embeddings. The official evaluation metric was used during this experiment, and it turned out that ELECTRA outperformed other pretrained language models. Using ELECTRA, both word embedding for each node in the graph, as well as document embedding for the whole tweet, are obtained. Each node was also annotated with the corresponding one-hot-encoded vector of its POS tag (45 features). Given that graph networks are ordering invariant w.r.t the nodes processed during

³Documentation about FLAIR, a popular open-source package for natural language processing that provides a range of state-of-the-art models for tasks such as named entity recognition, sentiment analysis, and part-of-speech tagging is available here: <https://github.com/flairNLP/flair>

the message-passing step, the order of words in the original tweet is lost. To maintain this information, each node is characterized with a feature vector of two dimensions that encodes the distance from the origin of each node (word) in the graph using sine and cosine positional encoding of a transformer model [412]. This vector is concatenated with the other node features.

6.3.3 Graph attention convolution and max pooling layer

Two GATV2Conv [60] layers are built into the model. The computation of dynamic attention scores characterizes this layer. Moreover, a multi-heads in the first layer (where the number of heads is set to four) is adopted because (as demonstrated previously by [414]) the learning process can benefit from employing multi-head attention and concatenating their outputs. As highlighted in Figure 6.1 the number of features used to represent each node is halved between the two layers. The output of the graph attention layer is a $2D$ matrix with the shape: Number of nodes (d) * Number of features (N). The maximum value is calculated along the dimension of size N reducing the dimension of the input tensor by one. As an example, consider the following matrix X .

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

Providing X as input to a 1D-max pooling layer returns the following array Y as output, where each y_i is computed taking the maximum of all the values along the i -th column of the matrix X .

$$Y = [y_1 \quad y_2 \quad y_3 \quad \dots \quad y_n]$$

6.3.4 Dense

The max pooling layer’s output is concatenated with the tweet embedding obtained from ELECTRA. This vector is fully connected to a dense layer, which is followed by a rectified linear unit function element-wise (e.g., $Relu(x) = max(0, x)$) and, finally, a dense layer with two units as output. These float values correspond to the softmax logits, a vector of raw (non-normalized) predictions.

6.4 Experimental setup

The chapter was submitted in *Subtask-1C (harmful tweet detection)* of CheckThat!’s Task 1 for the English language. Before addressing experimental setup and model training, the provided dataset is briefly described.

6.4.1 Dataset

The corpus includes a list of tweets labelled as harmful (1) or not (0). In addition, the **ID** of the tweets and their URLs are available. All samples are related to COVID-19. In general, a train, development and dev-test set are provided as well as an official test set that was used for the evaluation of the submissions. While for the first two dataset splits the gold labels were available, those of the official test set were held out till the end of the evaluation phase. The number of samples for all parts of the dataset is distributed as shown in table 6.1. The data was released as multiple tab-separated files (one per split)

Exploratory analysis shows the dataset imbalance with respect to the class labels. For the training set, the positive class samples correspond to only 8% of the total entries. Using the Tweet IDs provided, Twitter data were crawled via the official Twitter API ⁴. However, only a small subset of the samples (w.r.t. the training

⁴API Documentation available here: <https://developer.twitter.com/en/docs/twitter-api>

Dataset	Number of Samples	Label 0	Label 1
Train	3323	3031	292
Development	307	276	31
Dev-Test	910	828	82
Test	251	211	40

Table 6.1: Dataset statistics of all provided splits for English.

set, only 20% of the original tweets) was still available as the rest of this information had already been deleted by Twitter. Given this observation, it is chosen to discard the inclusion of social context information such as the number of tweet interactions (favourites, shares) and author features (follower following relationships, as well as user timeline tweets). Besides using the graph-based approach introduced in Section 6.1, further experiments on the dataset were performed featuring transformer-based methods as well as alternative graph construction techniques. Details on the results of these experiments are presented in Section 6.5.

6.4.2 Model training

The hyperparameter settings are in line with many of the decisions made in a study conducted in [432] and are used to fine-tune the proposed model subsequently. A Glorot uniform initializer [205] is selected to initialise the weights of the model. The model was compiled using binary cross entropy loss; this function calculates the loss with respect to two classes (i.e., 0 and 1) as defined in 2.7.

To improve the model performance and counteract class imbalance, class weights that correspond to a manual rescaling weight assigned to each class are chosen. Optimization is performed using the Adam optimizer [213]. To reduce overfitting, a learning rate scheduler reducing the learning rate by a factor of 0.9 with a step size of 25 epochs is adopted. The model architecture is depicted in figure 6.1, where the numbers of the various network hyperparameters are provided.

6.5 Results

6.5.1 Baseline

The organizers provide official baseline results based on random predictions. The metric to evaluate the task is the binary F1 score of the positive class label (harmful tweet). The official baseline result is 0.200 binary F1 on the evaluation test set. The baseline results are compared to the values of the proposed approaches achieved in Table 6.2.

An additional strong baseline based on a fine-tuned transformer model is also added. Different transformer architectures including BERT, RoBERTa and ELECTRA are evaluated regarding the classification task. It turned out that ELECTRA achieves the best results among these models (using 3 epochs for fine-tuning as recommended by [108]). The performance of this approach amounts to a 0.250 binary F1 score.

6.5.2 Proposed Approach

The main experiments are based on the model proposed in Section 6.3. Results are reported on the test set used for evaluation using the official binary F1 metric, as well as binary precision and binary recall of the positive class label.

As presented in Table 6.2 the submitted approach (*GCN+ELECTRA*) outperforms the official baseline by 8%. The official baseline approach generates class labels in random order.

Compared to the performance of the proposed baseline using ELECTRA, the *GCN+ELECTRA* outperforms this approach by 3%. An ELECTRA fine-tuning setup using 50 epochs is evaluated resulting in

Approach	Binary Precision	Binary Recall	Binary F1
Baseline	0.200	0.200	0.200
GCN+3-gram-ELECTRA	0.138	0.625	0.226
ELECTRA (3 epochs)	0.263	0.250	0.256
GCN+POS w/o word embeddings	0.166	0.650	0.264
ELECTRA (50 epochs)	0.275	0.275	0.275
GCN+ELECTRA	0.166	0.875	0.280

Table 6.2: Results (binary Precision, Recall and F1 of the positive class label) on the official test set for English with respect to different approaches.

almost as good of a performance as the finally submitted approach. However, the high number of epochs leads to strong overfitting on the training data.

In addition, results obtained by experimental setups that are evaluated as part of the development process of the submitted approach are reported. Table 6.2 *GCN+POS w/o word embeddings* refers to a setting where word embeddings are omitted and represent graph nodes by only considering one-hot encoded POS-tag vectors. In the *GCN+3-gram-ELECTRA* the model characterizes graph nodes by mean-pooled word embeddings of 3 subsequent words at each position. Thus, words' order that can be lost during graph convolution needed to be taken into account as well.

6.6 Discussion and future work

All other approaches used during the experiments result in lower performances when compared to the submitted model. This observation strengthens the decision of choosing to submit predictions generated with the model proposed in this chapter as it outperforms a range of other setups tested during the experiments. Obviously, a high recall score is reported, compared to low precision using the proposed model as it tends to frequently predict the positive class label, while it only predicts the negative one a few times.

As already mentioned above, the inclusion of social and user information in the model (as in [32]) can improve classification accuracy. By incorporating social features, the model is able to capture more contextual information that can be used to better understand the content and meaning of the text. For instance, knowing that a post has a high number of likes and retweets could indicate that it is particularly relevant or important to a specific audience, which in turn could be useful in predicting user behaviour or sentiment. Similarly, author profile information such as the number of followers could provide insights into their level of influence or expertise, which could be useful in predicting the credibility or reliability of the information being conveyed. Deleted tweets that can not be crawled anymore limit the usage of this information in the actual training set and their absence is probably one of the reasons for the poor performance of the model with respect to other submissions.

In future work, it would be interesting to analyze the contribution of such social context features. In addition, different types of embeddings (stacked or pooled) could be compared regarding the characterization of each node.

6.7 Conclusion

In this chapter, the approach to *Subtask- 1C at CheckThat!Lab@CLEF2022* based on graph data structures and transformer language models have been described. The proposed hybrid solution of a text graph and pretrained document embeddings should be studied in more detail as it leads to improvements over fine-tuning transformer-based models, as demonstrated on the given dataset. In addition, there is room left for including additional types of information (e.g. social media context) in these data structures which could be helpful in solving other tasks.

Chapter 7

Misinformation through images

This chapter proposes a framework to better understand the relationship between images and the truthfulness of the beliefs they convey, taking into account both objective and subjective concepts. The framework is supported by an annotated dataset, created through a pilot annotation study that evaluated the consistency and coherence of the proposed theoretical framework. The proposed framework and dataset are particularly relevant in the context of social media platforms, where images can have a significant impact on people's beliefs and opinions. By providing a more comprehensive understanding of visual content, the framework can help researchers recognise the factors that increase the spread of potentially misleading or false visual content.

7.1 Introduction

Social media have been considered the perfect environment for the spread of false beliefs and fake content, boosted by the use of Machine learning techniques that are designed to maximize engagement may have unintentionally resulted in the creation of a "*Frankenstein Monster*" that can disseminate fake news more effectively than real news [417]. In recent years, the role of recommender systems in shaping user preferences and opinions has been the subject of intense debate [19, 275]. On the one hand, these systems have been accused of contributing to the creation of echo chambers and filter bubbles, which can limit users' exposure to diverse perspectives and increase the spread of false content. On the other hand, it is essential to recognize that recommender systems can also play a critical role in slowing down the spread of false content. By tailoring recommendations to individual user's interests and preferences, these systems can help to surface high-quality, trustworthy content and reduce the visibility of false or misleading information. As such, a deeper understanding of how recommender systems can be used to combat the spread of false content is essential for addressing the challenges of misinformation and promoting a healthy information ecosystem, especially in presence of multimodal content, where a combination of texts and images are used together to convey messages. This thesis argues that while scholars debate the existence and impact of echo chambers and polarization effects caused by recommender systems, it is crucial to explore how these systems can mitigate the spread of false content.

Visual content recently took the lead as the preferred modality to share content leading to a shift from a text-centric to a visual-oriented experience on social media [230]. Visual content is becoming dominant on social media [230] because it has strong communicative potential due to its high perceptual immediacy. Still, it is not clear yet which forms of visual content may exacerbate (e.g., truthiness [284]), and even independently cause the spread of false beliefs, as well as which cognitive mechanisms are involved in the process [304].

While short and low-quality texts are common on text-based social media, e.g. tweets, have been shown to sometimes lack often enough context to be safely evaluated as truthful or misinformative, images are at

the same time richer and more perceptually immediate stimuli while still endowed with less context.

Images can provide a more complete and immersive experience for the viewer than a simple written description of the same scene. Additionally, visual content is often more engaging and attention-grabbing than text, making it more effective at conveying a message or story. Indeed, images are intrinsically ambiguous in many ways [193]. Thus images may foster the spread of false and biased beliefs without any image manipulation.

Observer-specific factors, such as their current opinion, may strongly affect their interpretation of an image [198, 445]. In addition, sophisticated ways to create and modify images have recently appeared, strongly affecting the weak link between images and reality. This is worrisome as even unrelated images appear to affect the interpretation of the truthfulness of content in other modalities [284]. Relying on a simplistic definition of “**truthful image**” may lead to the mischaracterization of the users and adoption of counterproductive support strategies [195]. Language and words are subjective and based on individual interpretation, which opens the door to deception. In contrast, pictures are often assumed to be more truthful and reflective of reality because the human brain perceives visual content as a direct representation of what exists. However, it is important to note that even images can be altered or manipulated, and thus may not always accurately reflect reality. Therefore, the dichotomy between true and false is insufficient to account for the multitude of combinations of different conditions and factors that could mislead the observers depending on their knowledge and the features of the image.

7.2 Motivation

Interestingly, while the impact of images combined with verbal content on the spread of false beliefs has been studied [71, 206], only limited work has been done on images in isolation. Understanding the link between image misinterpretation, propagation of false beliefs, users’ attitudes, and users’ image interpretation skills requires defining an accurate characterization of images, their interpretation and the resulting behaviours. Images can spread toxic beliefs [74, 355] still most attention in the scientific community as focused on text. In addition, sophisticated ways to create and modify images have been recently appearing [330, 338].

7.2.1 The role of computer-generated content

A crucial challenge related to video and image content is the so-called **DeepFakes** (a portmanteau of “*deep learning*” and “*fake*”) which refer to videos in which a person in an existing video or image is replaced with someone else’s likeness. They are characterized by hyper-realism, making depictions of people saying and doing things that never happened easily believable [68, 428]. While DeepFakes have the potential to revolutionize fields such as entertainment and marketing, they also pose significant risks to society. For example, DeepFakes can be used to spread misinformation and manipulate public opinion by creating fake news stories or altering the content of existing media [323]. Vaccari *et al.*, [407], found that DeepFakes can increase the uncertainty of people exposed to untrue visual information. Moreover, this resulting uncertainty reduces trust in news on social media. DeepFakes can also be used to harass or defame individuals by creating and distributing compromising or offensive material, such as the non-consensual production of pornography-related deepfakes used to substitute the victim’s face into a porn actor/actress’s body to create nonconsensual adult content [201]. Usually, the majority of the target are celebrities ¹. Many deepfake videos can be found on pornographic websites or underground communities, such as Telegram ². Some channels even offer (**illegally**) custom pornographic deep-fake services for a fee to interested users. Coupled with the speed at which information propagates in social media, deepfakes provide fertile ground for instances of defamation, blackmail, fraud, and threats to national security or democracy [12]. While Deepfakes also have a plethora of

¹Check Here the Report by TheSentinel.ai titled “Deepfakes 2020 The Tipping Point” <https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf>

²Check here: <https://www.bbc.com/news/world-60303769>, retrieved on April 7, 2023

constructive and legitimate uses in fields such as entertainment, healthcare and education, their destructive potential has indeed attracted sizeable attention, and the approaches for combatting said issues include changes in legislation and regulation, education and training, as well as the development of technology for detection [428].

Approaches to counteract threats like the previously mentioned DeepFakes include the usage of deep neural networks for the detection of artefacts resulting from the production of such content (for videos, see, for example, [53, 175, 197, 273, 386], for images see [76, 160, 185]). Such artefacts are, for example, related to image blending, the environment, behavioural anomalies, and audiovisual synchronization issues [267]. For example, the work of [53] relies on the complexity of heart rate dynamics derived from the facial video streams through remote photoplethysmography is vastly different between real and synthetic videos. In conclusion, there is a need to support young social media users in improving their ability to interact with image-based content that can bring potential risks and impacts for them. To provide effective support, it is necessary to understand better the different facets of the relationship between images and the truthfulness, rich interpretation, or beliefs they induce in the user.

7.2.2 Collaboration between publisher and social media to counteract visual misinformation

The relevance and urgency of promoting and also empowering the journalists' community let Reuters and Meta collaborate to create *The Reuters guide to Manipulated Media*³ that is a comprehensive resource developed in collaboration with the Facebook Journalism Project to help journalists and news organizations identify and address the growing problem of manipulated media. The guide provides a detailed overview of the various types of manipulated media, including deepfakes, synthetic media, and other forms of altered or fabricated content. It offers practical tips and strategies for verifying and debunking such content. It also includes a set of ethical guidelines for dealing with manipulated media, including the importance of transparency, accuracy, and fairness in reporting on such content.

Reuters highlight in its statement (reported below) all the multiple and complex facets that visual content brings and the urgency for multiple stakeholders to act together toward a better news environment populated by more aware players.

That's why we decided to partner with the Facebook Journalism Project to produce an in-depth course into manipulated media. It includes explanations and examples of new forms of synthetic media, along with an assessment of the full range of ways visual material may mislead. It also challenges participants to place themselves into a breaking news situation and consider the steps they can take to establish the facts around the pictures and videos they obtain.

7.3 Why true false dichotomy is not sufficient

The dichotomy of *true* or *false* can be defined as a binary approach to evaluating the veracity of the information based on the assumption that something is either objectively true or false. The dichotomy of true or false could not be sufficient for evaluating visual content as it fails to account for the nuanced and different levels of image interpretation and the various degrees to which an image can be altered. Techniques such as deepfakes or Photoshop can be used to create or alter visual content in a way that makes it appear authentic, making it difficult to determine the true nature of the content based on a simple true/false dichotomy. Another reason is that the concept of *truth* can be subjective and dependent on the context in which the visual content is presented. The same photograph may be perceived differently by different viewers based on their own experiences, beliefs, and biases, making it difficult to assign a universal truth value to the

³Source: <https://www.reuters.com/article/rpb-hazeldeepfakesblog-idUSKBN1YY14C>, retrieved on April 7, 2023

content. Furthermore, visual content often conveys more information and emotions than simple text, making it more powerful and influential. This means that the impact of visual content cannot be fully understood or evaluated without considering the context in which it is presented and the intended audience. Therefore, it is essential to consider a range of factors when evaluating the veracity of visual content, including the context in which it was created, the authenticity of the source, and the potential for manipulation. Simply relying on a dichotomy of true or false may not provide a comprehensive assessment of the reliability of the visual content.

Otherwise, relying on a simplistic definition of “*true image*” could lead to the mischaracterization of the users and the adoption of counterproductive educational and governance strategies. Thus, it is vital to distinguish image generation conditions and different levels of interpretation which lead an image to be naively considered false or true.

In other words, it is appropriate to enrich the characterization of images beyond the dichotomy of true-false because images bring significant ambiguity in generation conditions and interpretations, especially without context.

Multiple concepts relating to the true-false dichotomy are introduced to account for this ambiguity and the multiple factors that can affect the simplistic true/false classification. The need for this richer characterisation is also highlighted by Shen *et al.*, [363], that report how different characterisation levels impact the credibility estimation of a picture differently. These sets of conditions and interpretation levels correspond to different concepts and measures that are introduced informally through examples in the following paragraphs and later defined more formally in the subsequent sections.

7.4 Factors that contribute to images misinterpretation

7.4.1 Background knowledge and skills

A picture showing Superman (played by Christopher Reeve) and Ironman (by Robert Downey Jr.) fighting in the sky (as the one reported in Figure 7.1, generated using the *text-to-image* model created by OpenAi **DALLE-2**) would be false in many ways: **not authentic** as an official movie was not published with those two characters, **not truthful** as the scene was not recorded in reality, but it is the result of the combination of different pictures, and misleading for a significant part of the population, as, for example, someone may believe that such a kind of flying fight between fictional characters may be a part of movies’ scene.



Figure 7.1: An image of Superman and Ironman fighting in the sky above New York generated by the AI model DALLE-2.

On the other side, a user may consider another image (as the one reported in Figure 7.2 of Superman flying true because it is part of an official movie he watched (which was produced by a reliable and existing production company). In this case, one should say that the picture published by a reliable source is **authentic** instead of naively true. In this case, the term true would be ambiguous as the picture can be considered false in many ways, as described above and further discussed below. Once it is known that by true he meant **authentic**, this user should not be considered problematic (intended as someone who would benefit from educational activities) in terms of his ability to recognize that people don't fly and would not require support for this reason. Similarly, a non-problematic user may consider an image of Superman blocking a train naively **true** because not only is it a scene of that movie that he knows but also because he is informed that, in this case, the event of the actor playing Superman touching the front of that train actually happened (where the train was obviously still, or moving in reverse as in the old movie that used this trick that was surprising at the time). This picture is both **authentic and truthful** (instead of naively true) because it depicts a situation which physically happened, even if only from a static point of view, but can be still considered **misleading**.



Figure 7.2: Christopher Reeve playing Superman flying taken from the official movie

However, the image of Superman flying would not be true in this sense, and will not be simply truthful or false, as the sky and clouds were added in post-production. The event never physically happened with the visual elements present in the picture. Thus, if the user considered the image of Superman flying true in the sense of truthful, as some children could believe that Superman flies, he would be considered problematic and could benefit from educational intervention or support. This is another example of why more specific terms than true and false are needed when dealing with images.

7.4.2 Images manipulation

Concerning images' tempering, three types of image manipulation are introduced: **declared manipulation** when the publisher of the image declared (e.g. the making of) that it has undergone substantial manipulation (e.g. green screen), **detected manipulation** when the state of the art algorithms find a level of manipulation, and finally the **visible manipulation**, which describes if observers detect image manipulation, based on their own image analysis skills. While declared and detected manipulation are objective and can be assigned with an objective process, visible manipulation is subjective and depends on the observers.

Several other concepts are introduced that have a subjective nature, e.g. for some users, it may be easier or more difficult to process a picture and classify and estimate its authenticity and truthfulness without checking. For example, a painting may look similar to many other authentic paintings of Picasso, such as the one in Figure 7.4. However, an expert may recognize it as fake, as opposed to an authentic one, such as the portrait of Dora Maar in Figure 7.3. In this case, the fake painting has high **authenticatability** for ordinary observers but low authenticatability for a group of experts. Users believing that several pictures

with low authenticatability for ordinary observers are authentic would benefit from support and educational intervention. In practice, **authenticatability** corresponds to the subjective authenticity judgement of an observer when he cannot perform fact-checking on the image and has to rely on his own previous experience and skills for the judgement.



Figure 7.3: The portrait of Dora Maar is a 1937 oil on canvas painting by Pablo Picasso. It depicts Dora Maar, the painter's lover, seated on a chair.



Figure 7.4: A woman portrait in Picasso style generated by the AI DALLE-2

Similarly, for a picture that shows a face that an artificial intelligence model generated, as the one depicted in Figure 7.5, experts of that type of content may recognize the image's nature, which will have low credibility for them, but many observers will think it is an actually existing person, and it will have high credibility for common observers.



Figure 7.5: A GAN-generated image of a Politician

A low-quality photomontage or visible special effects will have low credibility for most observers. Users believing that several pictures with low credibility for ordinary observers are truthful may benefit from

support and educational intervention.

7.4.3 Subjective factors that influence interpretation

In practice, **credibility** corresponds to the subjective truthfulness judgement of an observer when he cannot perform fact-checking on the image and has to rely on his own previous experience and skills for the judgement.

Note that the definition of **authenticatability** and **credibility** focuses on the subjective experience of every single user or a supposedly uniform user group. An objective measure of authenticatability and credibility can be defined by defining a specific population, not necessarily uniform, and an aggregation procedure to combine the single subjective judgements of the population members (obtaining aggregated authenticatability and aggregated credibility). One can expect that if a significant part of the population evaluates the pictures, the aggregation will allow us to predict the response distribution of the rest of the population, similarly to election polls. However, the capabilities of image evaluation may enormously vary among groups. This process could still risk not appropriately covering groups sensitive to some pictures (e.g. fans of an unknown tv series).

Authenticatability and credibility focus on the correctness of specific attributes of the interpretation of an image, authenticity and truthfulness. However, an image usually induces a much richer representation in the observer's mind. This representation can comprise a putative event in the real world with its static attributes (e.g. location, objects), dynamic (e.g. objects movements) and also cognitive and social ones, such as identities, emotions, beliefs, and intentions, which are non-trivial to extract from a single picture as more context would be needed. However, the brain may intentionally or unintentionally extract them. In the following lines, the focus is on the subjective interpretation of these aspects.

Going back to comics, the picture of superman blocking the train may be considered naively true by a problematic user who interprets it as a picture from an event where the train was actually moving or pushing, and superman stopped it. As this fictitious event was just the result of some cinema trick that cannot be detected due to the lack of context in a single frame (e.g. the train may have been moving in the opposite direction with "superman" following it, and then the scene was played backwards), the frame must be considered to be misleading for this specific user, which is more precise than false, as it induced the user in believing in the event that never happened, where with "*belief in the occurrence of an event*" it is also considered the dynamic aspects (e.g. train movement direction) of the event as well as the cognitive and social ones, such as identities, emotions, beliefs, and intentions, which are non-trivial to extract from a single picture as more context would be needed.

7.4.4 Staged events

A user may consider naively true a picture with a soldier shooting a prisoner in a WW2 movie, thinking that the picture was taken during the real WW2. At the same time, it is the result of a staging process (as an example is reported a screen from the movie 1917, directed by Sam Mendes in Figure 7.6), even if in this case, the picture is truthful as people dressed like in the picture performed the observed action in the shown setting. However, the user's reasoning is still wrong because he interpreted it as a real WW2 event where a prisoner identical to the actor was killed. In this case, the image is considered misleading for this specific user, even if the action was not impossible, and something similar may have happened during WW2. Obviously, users believing in possible situations that are staged would be less problematic than those believing in impossible ones. However, the analysis and formalisation of the differences between these conditions and types of misleadingness, other than considering physical or cognitive/social aspects of the event, are left for future work.

Pictures from movies depicting actors in common situations, e.g. Clark Kent walking in the street or the characters of Friends laughing while sitting on a sofa in a pub, don't depict real common events but are staged by actors playing specific roles (acting), a fact that may not be recognised by users that don't know the



Figure 7.6: An image taken from the movie 1917, directed by Sam Mendes, that could be wrongly interpreted as a real photograph taken during a battle of WWI or WW2.

movie/series. Thus these pictures are also misleading for each member of this group that doesn't know the movies/series. If this group corresponds to the population of interest, then any aggregation of the individual misleadingness would result in the image being misleading for a significant part of the population. Otherwise, it will depend on the relative sizes of the population groups and the threshold adopted for significance.

7.4.5 3D reconstruction and beliefs misinterpretation contribute to misleadingness

Misleadingness (for a user), the fact that an observer will associate a false event to the image can result from several not directly observable factors. For example, it is difficult to understand from a picture if someone was self-defending or attacking somebody else or if a player intentionally touched the ball with the hand or if it just bounced on him during a football match. Observers automatically attribute intentions and beliefs to people in pictures, but they may fail due to lack of context or other limitations [363].

The classic tourist picture of holding the tower of Pisa may be misleading for people from remote regions of Kazakhstan, and the same would happen to Italians with a similar picture with the famous South Korean SunCruise hotel, designed to look like a cruise liner, as reported in Figure 7.7. This results from the lack of information necessary to reconstruct a 3d environment from a single 2d picture other than a lack of knowledge of this tradition.

Misleadingness may also result from missing social information such as identity, for example, pictures with doubles, impersonations or replicas, as in a typical picture with a fake Einstein, as some children may not know that he is dead.

7.5 A richer characterisation of image features

In the following sections, both objective and subjective features are operationalized. The concepts introduced through examples will be detailed in light of the annotation process presented later in the chapter.

Below there is a table that summarizes all the features introduced.

7.5.1 Authenticity

The proposed definition of authenticity relates closely to that adopted in image forensics and is not as strict as that of integrity. In general, an exhibit is authenticated when there is sufficient evidence that the exhibit



Figure 7.7: The Sun Cruise Resort&Yacht, located in Jeongdongjin on the east coast of South Korea, photographed from the beach.

Feature	Type	Description
Authenticity	Objective	Image is credited from a reliable source
Truthfulness	Objective	The image is a photograph of a real event
Declared Manipulation	Objective	Source declares image undergo a post-processing
Detected Manipulation	Objective	Models detect a form of manipulation
Authenticatability	Subjective	The received degree of authenticity
Credibility	Subjective	The perceived degree of truthfulness
Manipulation Visibility	Subjective	The perceived degree of manipulation
Uninformed Misleadingness	Subjective	The probability an image will mislead observed
Informed Misleadingness	Subjective	The probability an image will mislead observed, given truthfulness and authenticity of that image

Table 7.1: Table reports in each row an *Objective* or *Subjective* feature, and the third column provides a description of the feature.

is what the proponent claims it to be. Authenticity is not integrity. Authenticity means that the image is an accurate representation of the original event. Integrity refers to the information being unaltered from the time of acquisition until its final disposition. Several techniques can be adopted to trace the integrity of images through a three-level encryption algorithm such as the one proposed by [274] For instance, if to a JPEG image is applied compression and re-saved, the new image's authenticity is preserved, but the integrity is not. An image altered for fun or a bad photo altered to improve its appearance (but not its meaning) cannot be considered a forgery even though it has been altered from its original capture. As also reported by the official *Adobe Blog*

Content authenticity is when there is proper content attribution for creators and publishers, which helps ensure trust and transparency online. ⁴

Authenticity from a subjective perspective is what is perceived as authentic from an individual and could be perceived as inauthentic from another one. As highlighted by Fillitz *et al.*, [128], the tool of cultural relativism makes it reasonable that for two different people with different cultural backgrounds, the same picture could be perceived differently by them. Again [128] distinguished between two forms of authenticity in art:

1. **Nominal**: the correct identification of the origin, authorship or provenance of an object
2. **Expressive**: how much the work of art possesses the inherent authority of and about its subject

The proposed definition is the following: “An image is authentic when a reliable source published it, was credited with the authors and was not modified after publication.”

Also, if correctly reported, drawings, paintings, making-of pictures, and pictures with partial special effects and manipulations are authentic. The only source of subjectivity is the set of reliable sources. It does not rely on other concepts. To map sources into authenticity, fact-checking companies' methodology will be adapted. Those companies have to deal with a huge amount of information, so a selection process is needed to determine how to fact-check. Collaboration between fact-checkers and social media platforms is now established in the International Fact-Checking Network (IFCN) ⁵ that highlights the common traits of fact-checkers among their code of practices apart from being obviously non-partisan and set up for the purpose of fact-checking.

Reliable sources selection

Fact-checking companies are organizations that verify and validate the accuracy of the information disseminated to the public [157]. To do this effectively, these companies must select reliable information sources. Fact-checking companies use several criteria to determine the reliability of a source. One of the most important is the **credibility** of the source itself. This means the source has a track record of producing accurate and trustworthy information. For example, a source that has been in operation for a long time and has a reputation for being reliable is more likely to be considered reliable than a newer or less well-known source. Another factor that fact-checking companies consider is the **expertise** of the individuals or organizations that are providing the information. Suppose experts in a particular field are providing the information. In that case, it is more likely to be accurate than if it is being provided by individuals who are not experts in that field. For example, suppose a fact-checking company is looking to verify information about a medical topic. In that case, it is more likely to rely on sources such as medical journals or experts in the field of medicine than on sources that are not specialized in that area.

Moreover, suppose the information comes from a primary source, such as a government agency or a witness to an event. In that case, it is more likely to be reliable than if it is coming from a secondary source,

⁴Check Source here: <https://blog.adobe.com/en/publish/2021/10/22/content-authenticity-in-age-of-disinformation>, retrieved on April 7, 2023

⁵Check here <https://www.poynter.org/ifcn/>, retrieved on April 7, 2023

such as a news article that is based on information from another source. This is because primary sources are generally considered to be more reliable than secondary sources, as they are closer to the original source of information. Finally, fact-checking companies also consider the level of **bias** present in the information being provided. If the source has a clear bias towards a particular perspective, it is less likely to be considered reliable. Fact-checking companies strive to be as objective as possible and therefore prefer sources that present information in an unbiased manner. The Principles fact-checkers in the IFCN must abide by are summarized below, and those principles are adopted to qualify reliable sources by applying criteria for fact-checkers sources' eligibility and methodology:

- **Existence:** to be “legally registered as a company with the specific purpose of fact check”. In this case, this requirement and purpose are expanded to public institutions and news companies (newspapers, editors, publishers, broadcasters, and specialized blogs that adopt transparency as a standard for nominal authenticity). This is because the content of interest ““relates to or could have an impact on the welfare or well-being of individuals, the general public or society”” and is eligible because even ““if the organization receives funding from a local or foreign state or political sources, it should provide a statement on how it ensures its founders do not influence the findings of its reports””. (Principle 1)
- **Transparency:** each authentic source adopts as standard transparency of its sources: clearly, they ““must indicate the source of the facts they are checking”, and they “use primary sources wherever possible, always check a contestable source against other sources, do not unduly concentrate its fact-checking on any one side, and checks all key elements of claims against more than one named source of evidence””. (Principle 2). In this context, transparency means that the author/owner of the content is clearly credited.
- **Clear Methodology:** standards and methodologies they use to select, research, write and publish are clearly explained to users. In this case, specialized blogs with clear rules w.r.t. allowed content and its nominal authenticity are considered authentic. (Principle 4)

Examples of sources that can ensure authenticity to images, based on the notion of (*nominal*) authenticity, are listed below and are those who can ensure requirements in terms of authorship and post-processing:

- Press agencies: Associated Press, ANSA, LaPresse
- Media companies: Getty Images
- Publishers, Editors, broadcaster: New York Times, CNN, Netflix, BBC
- Public institutions: NASA

Not authentic images are the ones that came from:

- Media companies that do not satisfy requisites in crediting and post-processing documentation: IStock, Imgur
- Social Network: Reddit, FB, Instagram, even if the account is registered.
- General blogs with no references

7.5.2 Truthfulness

Truthfulness (and fakeness) applied to an image are strongly related to the fact that it was taken during a real event. Truthfulness refers only to a picture being related to a real event, certified by reliable sources. At the same time, it may not allow inferring the correct interpretation in terms of all the important aspects of the event during which it was taken. In this thesis, the adopted definition disregards intentionality but considers verifiability. Still, the manipulation of text or the creation of false text is an entirely different problem than the creation of false images. Thus truthfulness can be defined using source checking. A crucial

aspect of this definition is the concept of “*real event*”, which is not trivial as, in general, just the fact of taking a picture may affect the underlying event, especially when considering an extended definition of an event which involves the intention and mind state of the subjects in the picture. However, here is a simplified issue to enable a more immediate application and consider as real events those that can be photographed and don’t get substantial manipulation (e.g. no green screen). In addition, as content from fiction may often contain some parts created using methods like green screen and thus strongly detach from the correspondent event in the real world, scenes extracted from fiction will be considered truthful only when a reliable source reports that in those scenes no substantial manipulations were applied, while they can still being authentic even if the official image contains several levels of manipulation (e.g. all avatar pics would be authentic but few if any would truthful). **Truthfulness** does not focus on the interpretation of the image itself, as many different interpretations can be attributed to an image simply because some crucial aspects (such as intentions, e.g. Figure 7.8) are not visible, or elements are missing due to occlusions, or perspective related effects come into play. In other words, context may be missing in images in more ways than textual content.

The proposed definition of truthfulness is the following: “an image is truthful if it is authentic and the (reliable) publishing source certifies that it was taken during a real event in the real world and that no external elements were later added to it or removed from it (not composite, generated, or green screen) apart from publisher logo or watermark.” As a rule, **only authentic images can be considered truthful** because only reliable sources can ensure that the visual content is a photograph taken during a real event.

The picture will still be considered truthful if the source reports that the image underwent limited and standard pre-publication manipulation to improve its quality and understandability (e.g. improve contrast, brightness or zoom) but does **not** change the event depicted (e.g. rejuvenating an old actor to his early days or elements were added even if not central).

7.5.3 Declared Manipulation

Post-processing techniques are nowadays almost impossible to be correctly spotted. Moreover, post-processing is pervasive, given its meagre cost. The usage of post-processing techniques is playing a fundamental role and is becoming a distinctive aesthetic feature, too [57]. Various stakeholders in the visual content industry are limiting the allowed manipulation to avoid misinterpretation and encourage digital transparency. Photojournalist activity is becoming mediated by rigid codes of professional practice [350]. Getty Images is an American visual media company and a supplier of stock images, editorial photography ⁶, already prohibits manipulation of images for news and events but after the French law (see below) explicitly prohibits retouching the human body. Moreover, Associated Press, an American non-profit news agency headquartered in New York City ⁷, avoid the use of generic photos or video that could be mistaken for imagery photographed for the specific story at hand, or that could unfairly link people in the images to illicit activity. With respect to photos “No element should be digitally altered except as described below. Minor adjustments to photos are acceptable. These include cropping, dodging and burning, converting into grey-scale, eliminating dust on camera sensors and scratches on scanned negatives or scanned prints, and normal toning and colour adjustments. These should be limited to minimally necessary for precise and accurate reproduction and restoring the photograph’s authenticity. Changes in density, contrast, colour and saturation levels that substantially alter the original scene are unacceptable. Backgrounds should not be digitally blurred or eliminated by burning down or by aggressive toning. The removal of “red eye” from photographs is not permissible.” In France, to limit the usage of post-processing in body images that could boost not realistic and dangerous body stereotypes, the law imposes that any models appearing in commercial photography whose bodies have been made thinner or thicker by image processing software must be accompanied by the notice of *photographie retouchée* or retouched photograph. Failure to comply is punishable by a fine of more than \$44,000, or

⁶As defined by Wikipedia here https://en.wikipedia.org/wiki/Getty_Images

⁷Check Wikipedia here https://en.wikipedia.org/wiki/Associated_Press

30% of the money spent on advertising ⁸.

Manipulated content is in some way also allowed by established editors if it is properly documented and users can easily read and understand the disclaimer. Even if news sources are trying to cope with manipulated content also using disclaimers, they are not always effective. The efficacy of such countermeasures, especially in social media platforms, is still under debate. In McComb *et al.*, [256], 15 experimental studies were reviewed. Results showed that disclaimers are ineffective (and sometimes harmful) in the context of women exposed to thin body images with respect to their body image dissatisfaction, while Mena *et al.*, [260], highlight that w.r.t. fake news labelling content can be an effective way to reduce the spread of misinformation.

The proposed definition of declared manipulation is: “A declared manipulation is present when the established and reliable (authentic) content source points out to the reader in a clear way if and how the content is visually manipulated in terms of production (such as staging) or post-processing (such as blurring or splicing or green screen or CGI).”

7.5.4 Detected Manipulation

Manipulation and forgery detection using machine learning is a common task, and multiple methods are proposed in the literature to address the identifiability of manipulated content. Those methods are based on finding the region of inconsistencies with image features [322].

MantraNet, a pre-trained neural network, can spot forged regions. MantraNet is “an end-to-end image forgery detection and localization solution, which means it takes a testing image as input and predicts pixel-level forgery likelihood map as output” [437]. It comprises two sub-networks: the feature extractor and the local anomaly detection network. Different datasets were used for training, allowing the network to spot 385 different image manipulation types. Sun *et al.*, [385], (from Tencent Youtu Lab), motivated by the *state-of-the-art* performances of contrastive learning, propose a new framework for general facial forgery detection. Sadeghi *et al.*, [349], reviews the multimedia forensic algorithms proposed to detect forged images while [187] proposes to use a self-supervised method for detecting image manipulations. Their algorithm learns to detect and localize image manipulations (splices), even if it is not trained on manipulated images. Different architectures are reported to be effective in solving image forensic tasks. Gill *et al.*, [145], reviewed the state of the art of passive forgery detection approaches, also highlighting the limitations of this automation process:

1. the computational cost
2. the need to extend those methods also to audio and videos
3. the need for new methods that can distinguish between malicious forgery and just retouching like artistic manipulation

7.6 A crowd-sourced subjective characterisation of images

The **subjective aspect** of the proposed characterisation of images is reported in the following sections. As stated in Section 7.4 another aspect of interest is the subjectively perceived truthfulness and the potential to mislead, which can depend on the observer’s skills and knowledge. Users may be able to assess how likely a picture is related to an actual situation, is the result of a generative process or is being staged (for a movie or for misleading intentions, e.g. parade in support of a hated dictator). At the same time, users may be able to evaluate, with varying accuracy, if an image misleads other users. A straightforward way these features may be relevant for misinformation diffusion is through sarcastic behaviours where users may share

⁸Check here the text of the law (in French) https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000032411563/

a relatively misleading image, assuming that everyone could correctly interpret it. This may lead to the unintended spread of misinformation when this assumption is wrong. The goal is to overcome the simplistic dichotomy between true and false images to account for the multitude of combinations of different conditions and factors that could mislead the observers depending on their knowledge and the features of the image. Following this, the next Sections focus on the subjective experience of every single user.

7.6.1 Authenticatability

For an observer, the degree of **Authenticatability** of the image is how much he believes the picture to be certifiable based on his knowledge and skills. The Definition is provided:

“An image is authenticatable for a human observer if, based on his knowledge and image analysis skills, he believes it is Authentic, he believes that if he had the chance to do fact checking, he would find reliable sources that published and attributed the content.”

For a group of observers, the average aggregated Authenticatability is the mean of the degree of Authenticatability attributed by the observers in the group. Images are a rich information source and may induce a different sense of authenticity in various ways before fact-checking. When evaluating a user’s gullibility through images, it is important to consider not only his different reactions to authentic images and not authentic ones but also differentiate the reactions to very diverse images inside these two categories. Today, totally artificial images that are photorealistic and similar to authentic ones can be easily generated through computer-based methods. At the same time, many images can be authentic but be so unusual that they may not look authentic or may be intentionally created in that way, as some artwork looks like the result of a low-quality computer-based image alteration or generation process as well as painting style mixing Leonardo and Disney. Finally, some image processing (e.g. light balance) is the standard way of distributing some content (from weather news to fashion and back to stars, NASA’s ones). Authentic images can have low average Authenticatability and be or not misleading (e.g. they show something true that is still difficult to believe but do not push a different type of meaning). Authentic and credible images can still be misleading. Authentic images can have high average Authenticatability, be manipulated and still not be misleading.

7.6.2 Credibility

Tseng and Fogg, [404], affirm that image credibility is not an objective feature of the information itself but, ultimately, a perception of the user. They identify the **trustworthiness** and **perceived expertise** of the source of information as the most critical factors in forming credibility perceptions. People exposed to images usually take advantage of heuristics to assess the credibility of images by using the source’s reputation as a proxy for the content itself. Following this approach, the proposed definition is reported:

“An image is credible for a human observer if, based on his knowledge and image analysis skills, he believes it is true in a sense defined for Truthfulness, so the picture was taken during a real event for which reliable sources would be available if the observer had the chance to do fact-checking.”

For a group of observers, the average credibility is the mean of the degree of credibility attributed by the observers in the group.

Images are a rich information source and may induce a different sense of realism (thus truthfulness) in various ways before fact-checking. At the same time, many images can be truthful but be so unusual that they may look like the result of a low-quality computer-based image alteration or generation process. In practice, when an image has high average credibility for a group of expert users, even if it is false like the famous highly realistic picture of the Earth created for the iPhone background, and it has a high degree of credibility for a user, his gullibility may not be critical. Truthful images can have low credibility and be (or not) misleading (e.g. they show something true but still difficult to believe but do not push a different type of meaning). True and credible images can still be misleading. True images can be credible, manipulated and still not misleading.

7.6.3 Manipulation Visibility

The low cost of the digital post-processing tool and the relevance of images in our society (the use of photographic evidence in legal cases and the constant presence of social network platforms in our lives) affect the amount, and the relevance, of real and fake pictures that people see on social media [453]. Manipulated and forged images can affect individuals in various ways. In [287], researchers disagree about whether the low-level visual salience of objects in a scene, such as brightness or the high-level semantic meaning of the scene has the most influence on attentional allocation (and so the probability of detecting manipulation correctly) so an explicit way to account for this dimension apart from credibility and misleadingness is needed.

The proposed definition of **manipulation visibility** is: “For an observer, the degree of Manipulation of the image is how much he believes the post-processing is visible.”

For a group of observers, the average manipulation is the mean of the degree of manipulation attributed by the observers in the group. Chandakkar *et al.*, [74], compare the human ability to recognize image forger w.r.t algorithm performance. The authors examine the relationship between the saliency of an image and its effect on prediction performance. Using the eye-tracking data and the performance statistics from the evaluation test, develop an algorithm to predict the difficulty level of an image. People skills also affect the performance in the evaluation of forged images. Shen *et al.*, [363], hypothesize that People with greater levels of (a) photography and digital imaging experience and (b) general Internet skills will be more likely to perceive fake images as less credible than people with less experience/skills.

7.7 Misleadingness

Misleadingness is easy to attribute when the pictures are extracted from a fictional movie, but it becomes more tricky for truthful pictures of different origins. Going back to the acting and staging aspects, some political events may be staged to show the presence of high support for a leader that is actually hated and feared by the population, or empathy and compassion of the leader for the population, as well as the existence of good relationships between hostile political figures or countries.

Note that, like for credibility and autenticatability, the aggregated misleadingness is an objective measure once the population is defined. In contrast, the image interpretation of each member of the population is, obviously, member dependent and thus subjective. In fact, for naive observers, an objective difference exists between the actual cognitive-social aspects of the event and those of their interpretations of the images.

To attribute misleadingness to images with an objective process would require the following:

1. the often hard-to-access information on the real event conditions (e.g. presence of staging) but also
2. an objective (computational) process that predicts the cognitive-social aspects attributed by the different kinds of observers (e.g. children, naive part of the population) and
3. which kinds of observers form a relevant part of the population (considering, for example, how much they know about the topic)

A computational belief predicting process would require considering many factors, making its creation extremely challenging. The only other exact and objective alternative would be to sample the population, checking if their interpretation matches the real event description and aggregating the results. With a limited sample size or high similarity between individuals in the sample, this would risk missing groups that would be sensitive to some pictures, as discussed for credibility and authenticity.

Thus, while

1. authenticity and truthfulness can be assessed using a viable and objective procedure of fact-checking (once the reliable source are defined);
2. autenticatability and credibility can be assessed through a polling procedure;

3. for misleadingness other than a sampling procedure, an approximated and more subjective procedure may be adopted.

If information about the event is available (component 1), a group of informed assessors could estimate the interpretation errors of a broader population of observers (described by Component 3). For example, in the simplest case of a picture from a movie, the assessors would “only” have to consider if some people may not know that the picture is from a movie or cannot distinguish between movies and reality. In the case of political events or the WW2 picture, the assessors would have to consider if some people may not be aware of the staging process. Obviously, for generic pictures, disparate factors (e.g. knowledge of a specific cartoon, political belief, music band, diet, etc.) would determine the induced observers’ beliefs. So the expectations of each assessor will rely on his own specific opinion on the distribution of these factors in the observers’ population and his ability to analyze an image, hypothesizing possible interpretations and defining which factors will be crucial to process each image. A classical example of lacking context that may lead to misinterpretation is reported in Figure 7.8 where the picture shows a *"Hitler moustache"* inadvertently cast on the face of Angela Merkel (at the time Germany’s Chancellor) by the pointing finger of the Israeli Prime Minister ⁹.



Figure 7.8: The famous ‘Hitler moustache’ that casts a shadow over Merkel-Netanyahu meeting on February 25, 2014

A mismatch between assessed misleadingness for an assessor and the aggregation of misleadingness may be caused by factors depending on the sample or the assessor:

- Population sample’s limited size
- Population sample’s distribution not matching population one (biased sample);
- Assessor’s limited understanding of the population distribution;
- Assessor’s limited understanding of the picture peculiarities.

Ultimately, the assessed misleadingness would be measured as the aggregation of the estimated misleadingness provided by the single assessors. Various procedures could be used for the aggregation depending also on the format of predictions of the assessors. If the assessors provide the percentage of the population that will be misled, then the aggregation can be the median value. Other approaches can be used to take into account the shape of the distribution, which may also be informative on the strength of the misleadingness of the picture and the peculiarities of the social groups it may affect.

⁹Source: <https://www.france24.com/en/20140226-merkel-netanyahu-hitler-moustache-photo>, retrieved on April 7, 2023

The information about the real event conditions during which the picture was captured, component (I) for misleadingness definition, deserves particular attention because, in most cases, when observing a picture it may not be available. The process of predicting the misleadingness of a picture without context information, which is defined as estimating the **uninformed misleadingness** of the picture, is more common than expected and often takes place when users decide to share pictures on social media.

In the absence of this information (I), assessors would have to rely on their different knowledge and intuitions to evaluate if other observers would be misled other than their different assumptions on the observers' population. For some images, the lack of information could create a strong variability between the assessors' responses. If, for an image, such variability is substantially higher than for the responses obtained when the information is available (**informed misleadingness**), then processing such an image would be particularly difficult for several users and require relevant knowledge on the topic of the image. Actually, the difference between uninformed and informed misleadingness can characterise the impact of contextual information necessary to interpret an image.

When ground truth information is not available to the assessors, a staged picture of Mike Tyson fighting Mohammed Ali may be trustworthy for assessors that don't know much about boxing as they may believe it represents a real match as many others. Nobody will be brought into believing something that did not exist. Instead, assessors that know about boxing will consider the image not trustworthy as they know that there was never a real fight between them, but other common observers may fall for it.

Note that for **uninformed misleadingness** and for sampling-based evaluation of authenticatability, credibility, and the users are not informed about the real nature of the picture but have to rely on their own knowledge and observation skills. On the other hand, assessors are informed about images' nature (authenticity and truthfulness) when they provide an estimate of **informed misleadingness**. In conclusion, **Informed misleadingness** refers to the situation where an image is intentionally or unintentionally misleading. Still, the observer is aware of the context (in this case, the context is intended as the objective feature *authenticity* and *truthfulness* who is aware of) or has the necessary information to interpret the image accurately. On the other hand, **uninformed misleadingness** refers to the situation when the observer does not have the contextual information. These concepts highlight the importance of context and accurate information when framing the composite concept of truthfulness related to visual content. It is also important to consider the potential biases and preconceptions of the viewer, as these can influence how an image is interpreted. To do so, a set of more subjective concepts have been introduced, framing misleadingness as a two-stage procedure to reduce variability due to annotators' knowledge and background.

7.7.1 Uninformed Misleadingness

The popularization of digital techniques to modify images has lowered the trustworthiness of visual content because the capability to imply a matter of fact and the truth is reduced [186]. Low-cost techniques reduce the possibility that an image reflects and is based on evidence to believe, which is a base for trust [442]. Grandison, [156] defines trust in internet applications as ““the firm belief in the competence of an entity to act dependably, securely and reliably within a specified context (assuming dependability covers reliability and timeliness)””. In this setting, the concept of misleadingness is introduced to understand the factor that affects images' trustworthiness. It evaluates how complex a user thinks it is to interpret an image for others when he does not actually know additional information about it, so he does not have access to more context and cannot perform fact-checking. This difficulty in interpreting an image is closely related to the potential misleading of the image, and the evaluation of misleadingness is based on the metacognitive skills of the judge.

To reduce uncertainty and the spreading of false information, Karduni *et al.*, [202] proposed a framework that serves as a visual analytic system to support the investigation of misinformation on social media, a visual interface that presents features related to real vs suspicious news thus raising awareness of multiple features that can inform the evaluation of news account veracity.

The proposed definition of uninformed misleadingness is the following:

An image is Misleading if it will make people believe that a real event never happened or vice versa that something that never happened is actually happening

Suppose substantial knowledge is necessary to determine the correct interpretation of a picture. In that case, the annotators, without additional information, may mislead themselves and assume that the image is not misleading, so there is significant variability in the evaluation because of (i) the absence of context (such as a caption) that could provide more information and (ii) the users' specific knowledge and expertise. In this case, without context, the background knowledge and skills of the annotators can be considered confounding factors and a step to reduce the uncertainty generated is needed.

7.7.2 Informed Misleadingness

The information context of images in social media is not always absent. A minimal fact-check with a heuristic contributes to the interpretation, so a two-stage approach is adopted. The second step is equipped with additional information that can reduce the uncertainty about the images' interpretation, and the resulting misleadingness evaluation can be more accurate and less variable. **Informed misleadingness** is equal to the uninformed one except that objective features, such as **truthfulness** and **authenticity**, are provided to the annotator. This knowledge inoculation that characterizes the second step of the misleadingness annotation help to understand how much is essential the objective features provided are. This approach is similar to the one of Mena *et al.*, [260], where results show the positive effects of flagging fake news.

7.8 Dataset Annotation procedure

After selecting a set of reliable (or not) sources, and defining objective and subjective features, the first step is downloading images with the corresponding URL and assigned topic. After fact-checking, the objective features were annotated. For authentic images, when possible, a caption that describes the image is stored along with the corresponding image. In the figure below, all the steps of dataset annotation are reported.

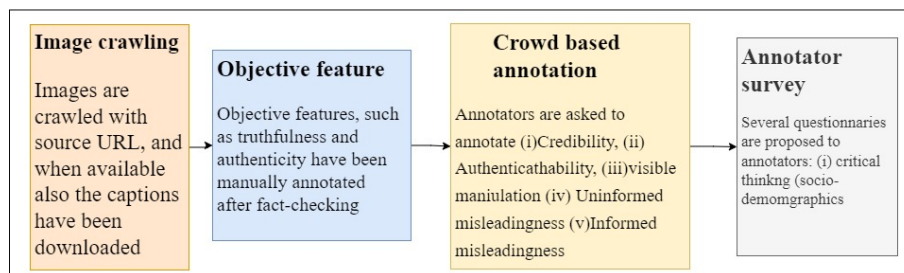



Figure 7.9: Flowchart of the dataset annotation process.

Once objective features have been annotated, the images are uploaded in Qualtrics software, where annotators can access remotely and annotate images.

Annotators were recruited with the Sona System instance of Bicocca University. A total of 53 annotators took part in the study in exchange for 0.2 credit for around 1 hour of work. Once completed the registration and consent form annotators start the task. Each image was shown twice. The first-time annotators answer to credibility, authenticity, manipulation visibility, and uninformed misleadingness. For a subsample (10%) of authentic images, annotators provide a brief description of the image (captioning). Captioning can be used to evaluate the correspondence between official captions and the ones provided by annotators. Metrics that are suited for this task are: BLEU [334], but its main limitation is that it does not consider the meaning of words and it looks only for exact word matches; Hessel *et al.*, proposed a *reference-free* evaluation

metric, named CLIPScore [177]. In Kilickaya *et al.* an extensive evaluation of the existing image captioning metrics is provided [209].

In riferimento all'immagine qui sotto, ti chiediamo di indicare su una scala da 1 a 5 (1 = Per niente, 5 = Moltissimo) quanto ritieni che l'immagine sia...



Per niente 1 2 3 4 Moltissimo 5

... stata pubblicata da un editore affidabile, che ne riporti correttamente il contenuto e l'autore

... una foto che ritrae un evento reale

... stata soggetta a una manipolazione grafica

Ingannevole (nel senso che farà credere che un evento reale non sia mai accaduto o viceversa che qualcosa di mai accaduto sia realmente accaduto)
Nota bene: ingannevole non significa falso, una immagine può essere ingannevole indipendentemente dal fatto che sia o meno una fotografia.

Figure 7.10: A screen of the Qualtrics annotators page.

The second time the image is shown, decorated with truthfulness and authenticity labels, the annotators are asked for informed misleadingness. After the annotation, participants answer multiple questionnaires:

- **Socio-Demographic:** Age, gender, level of education, political orientation
- **Critical thinking disposition scale (CTDS)** [370]: is a tool used to measure an individual's dispositions towards critical thinking. It is designed to assess the extent to which an individual habitually engages in critical thinking in their everyday life. The CTDS consists of a series of statements that individuals respond to on a Likert scale, with options ranging from "*strongly disagree*" to "*strongly agree*".
- **Bullshit receptivity scale:** [312]: is a tool used to measure an individual's susceptibility to accepting statements or claims that are vague, ambiguous, or lack evidence. It is designed to assess the extent to which an individual is likely to be influenced by statements that are not supported by logic or evidence. The scale gives a better understanding of the underlying cognitive and social mechanisms determining if and when bullshit is detected.

7.8.1 Dataset Description

The dataset is expected to have some similarities with images in social media concerning sources and topics. The sources are selected between the ones listed in Section 7.5.1 and differ in terms of authenticity and truthfulness. The dataset covers the following topics:

1. **Environment and Natural Disaster:** include images about floods, both real and GAN-generated, or images from professional photographers of storms and lightning.
2. **Politics:** images of politicians during official events or GAN generated ¹⁰
3. **Commons:** common or strange objects.

¹⁰The GAN generated images are taken from here: <https://thisxdoesnotexist.com/>

topic	truthfulness	authenticity	samples
commons	F	F	48
commons	F	T	18
commons	T	T	58
naturaldisaster	F	F	27
naturaldisaster	F	T	28
naturaldisaster	T	T	74
politics	F	F	39
politics	F	T	23
politics	T	T	91
space	F	F	35
space	F	T	26
space	T	T	55

Table 7.2: The table reports the sample size for each topic and is also divided between truthfulness and authenticity.

4. **Space:** images from NASA taken from missions (Hubble, Apollo), posters celebrating different missions, and staged scenes from different movies or forged images.

Images sometimes include people pictured in them. The table below reports the sample size divided by topic, authenticity and truthfulness.

7.9 Annotation results

A total of 27 annotators completed the task correctly. Annotations show that:

- **Informed misleadingness** decreases for **truthful and authentic (TT)** images, while increases for the other two categories (**Authentic & non truthful (TF)**, and **Not truthful & not authentic (FF)**)
- Average (uninformed) misleadingness is higher for not authentic images: $FF > FT > TT$
- Annotator’s variables, such as political orientation and bullshit receptivity scales, seem not to influence annotation, but the low sample size does not allow generalisation.
- **Uninformed Misleadingness** correlates more with manipulation visibility than credibility.

In the table below, all the features’ correlations are reported. **Credibility** and **authenticability** have the highest value for Pearson correlation. Uninformed misleadingness is moderately correlated with manipulation visibility and negatively correlated with credibility.

Figure 7.12 below reported the aggregated visualization in polar coordinates of all images’ features. It appears that truthful images have, on average, higher credibility and less manipulation visibility, while not truthful images are characterized by higher misleadingness.

Below the annotations are analyzed, grouping annotators’ and images’ features. Annotators are divided into two groups: (i) Above the average of Critical thinking (CTDS) and (ii) below. Images are divided into two groups: (i) above the average credibility and (ii) below. Some differences in the annotation emerge:

- For Annotators above the average CTDS n images that are not truthful (F) but above the average credibility, the misleadingness difference is 1.4, higher than the same images for annotators below 0.8.
- For Annotators below the average CTDS, truthful images (T) but below the average credibility, misleadingness difference decreases between uninformed and informed misleadingness annotation stage.

Correlation					
		autent	credib	visibility	Mislead
autent	Correlazione di Pearson	1	,712**	-,241	,057
	Sign. (a due code)		<,001	,226	,779
	N	27	27	27	27
credib	Correlazione di Pearson	,712**	1	-,334	-,020
	Sign. (a due code)	<,001		,088	,923
	N	27	27	27	27
visibility	Correlazione di Pearson	-,241	-,334	1	,547**
	Sign. (a due code)	,226	,088		,003
	N	27	27	27	27
Mislead	Correlazione di Pearson	,057	-,020	,547**	1
	Sign. (a due code)	,779	,923	,003	
	N	27	27	27	27

** . La correlazione è significativa a livello 0,01 (a due code).

Figure 7.11: The table reports the correlation between images features

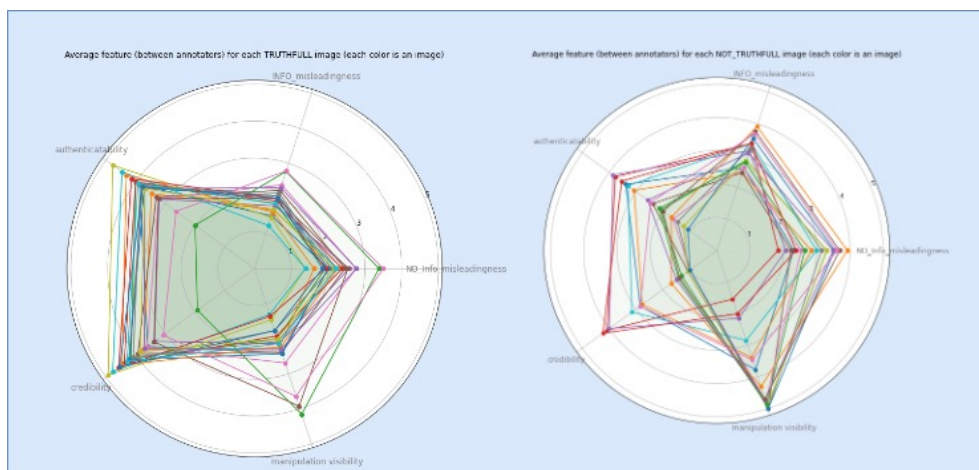


Figure 7.12: Polar graph of truthful (Left) and not truthful images (Right). Each pole represents an annotated feature.

- Images that are truthful (T) but below the average credibility have a much higher decrease in terms of misleadingness for annotators below the average critical thinking threshold.

crit_think_group	truthfulness	above_avg_cred	Uninfo	Mislead	Info_mislead	avg_diff_misl
high	F	high	2.35		3.75	1.40
	F	low	2.83		2.97	0.14
	T	high	1.82		1.58	-0.24
	T	low	3.70		3.50	-0.20
low	F	high	2.34		3.16	0.81
	F	low	3.35		3.23	-0.12
	T	high	2.44		2.24	-0.20
	T	low	3.38		2.75	-0.62

Table 7.3: The table reports uninformed and informed misleadingness for images divided into Above (*High*) and below (*Low*) the average credibility while annotators are divided into above (*High*) and below (*Low*) the average critical thinking.

These results suggest that annotators above the average CTDS are more sensitive to misleading content, while annotators below the average tend to give minor rates to images. This can originate from the fact that there is less awareness in the latter group of annotators.

7.10 Conclusion

This chapter presented a framework to contextualize and map images with low or absent context with a mix of objective (verified through fact-checking) and subjective (crowdsourced using annotators) features.

The images crawled are typical of social media and mix all the objective features introduced.

The framework has been tested with a sample of annotators that completed the annotation of 50 images. Overall, the proposed subjective features seem to be coherent between them and provide a helpful tool to map images without context with respect to their truthfulness and potential to mislead users.

It emerges that uninformed misleadingness correlates with manipulation visibility. At the same time, annotators' variables (in particular Critical Thinking) affect annotations because annotators with higher CTDS seem more sensible and tend to assign higher misleadingness to images.

7.10.1 Future work

The annotated dataset will be used in free-interaction browsing activities. Participants will be exposed to a sequence of grids of images. In each one, participants can select one and see it in full-screen mode. Using the annotated dataset to compose the grids, the goal will be to understand only from participants' selection if they could benefit from educational activity.

Chapter 8

Conclusions

Digital platforms centred on user-generated content like social media, currently rule the information and entertainment industries everywhere. Projections up to 2027 indicate a clear rise in the proportion of users who are engaged on these platforms. In order to reach a consensus and achieve widespread adoption of a policy framework, it is essential to comprehend the results of this collective involvement and its impacts on individual well-being and welfare. In the coming years, a transformative regulatory innovation will be necessary to protect society and fully utilize social media's advantages for individuals, organizations, and private businesses.

This thesis addressed the issues of social media threats from different points of view. First, the content can be classified and labelled with machine learning models that can serve as detectors in social media. Fake news, cyberbullying and harassment, in fact, can take many multimodal forms and have serious consequences for the mental and emotional well-being of those targeted. Additionally, social media platforms can be used to spread misinformation and propaganda, which can lead to confusion and mistrust among users. Social media can also contribute to developing body image and self-esteem issues, as users are exposed to an endless stream of highly curated and often unrealistic images of other people's lives.

In the proposed architecture of the Social media companion, given the importance of these threats, ML models can be developed with the goal of building a detection system to spot these threats as proposed in Chapter 6 with promising results, even if limitations of detecting and labelling content are abundant both from the side of model training and the effectiveness of these labels in educating users. The proposed model ingests texts to label harmful tweets relate to COVID-19 but its main feature derives from the application of graph networks that allow to include also social information such as metadata about the author of the tweet or the number of likes or reshares. Methods for addressing threats have been proposed and the integration of support for multimodal content and otherwise invisible network-based threats directly into the user feed can help users to discriminate and have a more comprehensive understanding of what the piece of text means. However, it remains an open question to which extent the analysis and visualization of the content lead to more threat awareness among users of social media platforms.

Social media users are another crucial aspect, especially teenagers, which should be better educated through digital media literacy activities, which are gaining momentum to counteract social media threats and protect individuals from misinformation, cyberbullying, and other negative effects of digital media. It can, in fact, help users to identify and correctly respond to potential threats. Digital media literacy can empower users to take control of their online experience by teaching them how to set privacy settings and control the content they see. It can also encourage them to think critically about the impact of social media on their mental and emotional well-being and to develop healthy habits for using these platforms. One way to improve digital media literacy is through game-based activities. Using a game-based approach, individuals can engagingly learn about digital media, and positive results have been shown in Chapter 5. The proposed structure of the activity, composed of an educational talk and a game base activity has proven

to increase, with statistically significant results, at least in the short term, the perceived influence of social media. Game-based structures can also provide a safe and controlled environment for individuals to practice identifying and addressing digital media threats. The main limitation of these educational activities is the fact that is unclear if, in the medium and long term, they are still effective.

The information that circulates on social media is becoming more complex to interpret safely because of multimodality and the lack of context, given also the shortening of pieces of information and the overabundance of items that reduce the attention budget that can be dedicated to each one. Visual content, in particular, is taking the lead and analysis to understand better the factors that can increase or decrease images' misleading potential, which has been introduced and tested in Chapter 7. This characterisation is innovative because it considers objective and subjective features and is purposely designed to deal with the interpretation level taking advantage of different information steps. It was found that different levels of critical thinking (CTDS) affect the annotations this result suggests that annotators above the average CTDS are more sensitive to misleading content, while annotators below the average tend to give minor rates to images. This can originate from the fact that there is less awareness in the latter group of annotators. It was also found that Uninforme misleadingness positively correlates with statistical significance with manipulation visibility. From a macro perspective, improving AI-human value alignment in recommender systems is crucial for creating a more inclusive and fair system for all stakeholders. However, operationalizing collective-well-being metrics and creating a pipeline to enforce and dynamically update those metrics is a challenging task from a technical perspective. This requires a deep understanding of the underlying algorithms and integrating ethical considerations into the design and development process.

Here, an approach based on educationally-managed social media communities has been proposed in 3. Moreover, gathering the necessary data and resources to measure and improve collective well-being may be difficult. Despite these challenges, it is important to continue working towards improving AI-human value alignment in recommender systems to ensure that they benefit all users and society as a whole.

A lot of work is still needed to create a safer and well-being-oriented social media environment. This work must be interdisciplinary and should go in the direction of a disruptive regulatory framework that must become more inclusive for all the stakeholders involved and a more transparent algorithmic auditing process.

In conclusion, this thesis has highlighted the multiple problems and multidisciplinary nature of addressing social media threats. The research conducted and the approaches adopted have drawn from a variety of fields, mainly computer science but also psychology. The diversity in perspectives and methodologies has allowed for a more comprehensive understanding of the problems. The importance of addressing these issues cannot be understated, as they have significant ramifications for individuals, communities, and society.

Bibliography

- [1] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.
- [2] H. Abdollahpouri and R. Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness. *CoRR*, abs/1907.13158, 2019.
- [3] H. Abdollahpouri, R. Burke, and B. Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In *The thirty-second international flairs conference*, 2019.
- [4] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.
- [5] D. Abel, J. MacGlashan, and M. L. Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.
- [6] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.
- [7] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang. Reducing recommender system biases: An investigation of rating display designs. *MIS Quarterly*, 43(4):1321–1341, 2019.
- [8] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattacharyya. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, 139:112851, 2020.
- [9] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.
- [10] K. O. Akomeah, U. Kruschwitz, and B. Ludwig. University of regensburg @ pan: Profiling hate speech spreaders on twitter. In *Proceedings of the 12th Conference and Labs of the Evaluation Forum (CLEF2021)*, pages 2083–2089. CEUR Workshop Proceedings (CEUR-WS.org), 2021.
- [11] K. O. Akomeah, U. Kruschwitz, and B. Ludwig. Ur@nlp_a_team @ germeval 2021: Ensemble-based classification of toxic, engaging and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 95–99, Duesseldorf, Germany, sep 2021. Association for Computational Linguistics.
- [12] M. Albahar and J. Almalki. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22):3242–3250, 2019.

- [13] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
- [14] H. Alcaraz-Herrera, J. Cartlidge, Z. Toumpakari, M. Western, and I. Palomares. Evorecsys: Evolutionary framework for health and well-being recommender systems. *User Modeling and User-Adapted Interaction*, pages 1–39, 2022.
- [15] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, 2020.
- [16] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.
- [17] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021.
- [18] M. B. Almourad, J. McAlaney, T. Skinner, M. Pleya, and R. Ali. Defining digital addiction: Key features from the literature. *Psihologija*, 53(3):237–253, 2020.
- [19] S. Altay, M. Berriche, and A. Acerbi. Misinformation on misinformation: Conceptual and methodological challenges. *Social Media+ Society*, 9(1):20563051221150412, 2023.
- [20] A. Alvarez-Socorro, G. Herrera-Almarza, and L. González-Díaz. Eigencentality based on dissimilarity measures reveals central nodes in complex networks. *Scientific reports*, 5(1):1–10, 2015.
- [21] I. Amarasinghe, D. Hernández-Leo, and A. Jonsson. Data-informed design parameters for adaptive collaborative scripting in across-spaces learning situations. *User Model. User-Adap.*, 29(4):869–892, 2019.
- [22] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [23] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, pages 2155–2165, 2020.
- [24] M. Andrejevic. Public service media utilities: Rethinking search engines and social networking as public goods. *Media International Australia*, 146(1):123–132, 2013.
- [25] C. Andris, D. Lee, M. J. Hamilton, M. Martino, C. E. Gunning, and J. A. Selden. The rise of partisanship and super-cooperators in the us house of representatives. *PloS one*, 10(4):e0123507, 2015.
- [26] S. Arnstein. “a ladder of citizen participation”: Journal of the american institute of planners (1969). In *The City Reader*, pages 290–302. Routledge, 2020.
- [27] Avaaz. Facebook’s algorithm: A major threat to public health, 2020.
- [28] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644, 2011.
- [29] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

- [30] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- [31] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- [32] V. Balakrishnan, S. Khan, and H. R. Arabnia. Improving cyberbullying detection using twitter users’ psychological features and machine learning. *Computers & Security*, 90:101710, 2020.
- [33] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrent social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, 2013.
- [34] J. Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34, 2021.
- [35] J. Bandy and N. Diakopoulos. # tulsaflop: A case study of algorithmically-influenced collective action on tiktok. *arXiv preprint arXiv:2012.07716*, 2020.
- [36] J. Bandy and N. Diakopoulos. More accounts, fewer links: How algorithmic curation impacts media exposure in twitter timelines. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–28, 2021.
- [37] S. Banker and S. Khetani. Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, 38(4):500–515, 2019.
- [38] S. Barocas and A. D. Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [39] D. Bawden and L. Robinson. Information overload: An overview, 2020.
- [40] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*, 2018.
- [41] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. In *WebDB*, 2009.
- [42] J. Becker, D. Brackbill, and D. Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076, 2017.
- [43] J. Becker, E. Porter, and D. Centola. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722, 2019.
- [44] P. L. Beed, E. M. Hawkins, and C. M. Roller. Moving learners toward independence: The power of scaffolded instruction. *The Reading Teacher*, 44(9):648–655, 1991.
- [45] R. M. Bell and Y. Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 43–52. IEEE, 2007.
- [46] A. Bellogín, P. Castells, and I. Cantador. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, 20(6):606–634, 2017.
- [47] A. Bessi. Personality traits and echo chambers on facebook. *Computers in Human Behavior*, 65:319–324, 2016.

- [48] A. Bessi and E. Ferrara. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7), 2016.
- [49] I. Beyens, E. Frison, and S. Eggermont. “i don’t want to miss a thing”: Adolescents’ fear of missing out and its relationship to adolescents’ social needs, facebook use, and facebook related stress. *Computers in Human Behavior*, 64:1–8, 2016.
- [50] S. Bird and E. Loper. Nltk: the natural language toolkit, 2004.
- [51] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.
- [52] C. Blex and T. Yasseri. Positive algorithmic bias cannot stop fragmentation in homophilic networks. *The Journal of Mathematical Sociology*, 46(1):80–97, 2022.
- [53] G. Boccignone, S. Bursic, V. Cuculo, A. D’Amelio, G. Grossi, R. Lanzarotti, and S. Patania. Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In *International Conference on Image Analysis and Processing*, pages 186–195. Springer, 2022.
- [54] M. Boeker and A. Urman. An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022*, pages 2298–2309, 2022.
- [55] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287, 2002.
- [56] R. Borges and K. Stefanidis. *On measuring popularity bias in collaborative filtering data*. CEUR-WS.org, 2020.
- [57] E. Borges-Rey. News images on instagram: The paradox of authenticity in hyperreal photo reportage. *Digital Journalism*, 3(4):571–593, 2015.
- [58] E. Bozdog and J. van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, 2015.
- [59] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.
- [60] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.
- [61] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- [62] M. V. Bronstein, G. Pennycook, A. Bear, D. G. Rand, and T. D. Cannon. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of applied research in memory and cognition*, 8(1):108–117, 2019.
- [63] A. J. Brown. “should i stay or should i leave?”: Exploring (dis) continued facebook use after the cambridge analytica scandal. *Social media+ society*, 6(1):2056305120913884, 2020.
- [64] A. Bruns. After the ‘apocalypse’: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566, 2019.

- [65] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.
- [66] R. D. Burke, H. Abdollahpouri, B. Mobasher, and T. Gupta. Towards multi-stakeholder utility evaluation of recommender systems. *UMAP (Extended Proceedings)*, 750, 2016.
- [67] A. L. Burrow and N. Rainone. How many likes did i get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology*, 69:232–236, 2017.
- [68] S. Bursic, A. D’Amelio, M. Granato, G. Grossi, and R. Lanzarotti. A quantitative evaluation framework of video de-identification methods. In *2020 25th international conference on pattern recognition (ICPR)*, pages 6089–6095. IEEE, 2021.
- [69] C. Cadwalladr and E. Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17:22, 2018.
- [70] M. Cannon, S. Connolly, and R. Parry. Media literacy, curriculum and the rights of the child. *Discourse: Studies in the Cultural Politics of Education*, 43(2):322–334, 2022.
- [71] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161, 2020.
- [72] S. Casale and G. Fioravanti. Factor structure and psychometric properties of the italian version of the fear of missing out scale in emerging adults and adolescents. *Addictive behaviors*, 102, 2020.
- [73] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi. Artificial intelligence and the ‘good society’: the us, eu, and uk approach. *Science and engineering ethics*, 24(2):505–528, 2018.
- [74] P. S. Chandakkar and B. Li. Investigating human factors in image forgery detection. In *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia*, pages 41–44, 2014.
- [75] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.
- [76] X. Chang, J. Wu, T. Yang, and G. Feng. Deepfake face image detection based on improved vgg convolutional neural network. In *2020 39th Chinese control conference (CCC)*, pages 7252–7256. IEEE, 2020.
- [77] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*, 2020.
- [78] L. Chen et al. Building a profile of subjective well-being for social media users. *PloS one*, 12(11), 2017.
- [79] M. Chen, Y. Wang, C. Xu, Y. Le, M. Sharma, L. Richardson, S.-L. Wu, and E. Chi. Values of user exploration in recommender systems. In *Fifteenth ACM Conference on Recommender Systems*, pages 85–95, 2021.
- [80] R. Chen, Q. Hua, Y.-S. Chang, B. Wang, L. Zhang, and X. Kong. A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access*, 6:64301–64320, 2018.
- [81] X. Chen, P. Tsaparas, J. Lijffijt, and T. De Bie. Opinion dynamics with backfire effect and biased assimilation. *PloS one*, 16(9):e0256922, 2021.

- [82] Y.-W. Cheung and K. S. Lai. Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280, 1995.
- [83] C.-W. Chiang and M. Yin. Exploring the effects of machine learning literacy interventions on laypeople’s reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.
- [84] T. Choolarb, J. Premsmith, and P. Wannapiroon. Imagineering gamification using interactive augmented reality to develop digital literacy skills. In *Proceedings of the 2019 The 3rd International Conference on Digital Technology in Education*, pages 39–43, 2019.
- [85] T. H. H. Chua and L. Chang. Follow me and like my beautiful selfies: Singapore teenage girls’ engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior*, 55:190–197, 2016.
- [86] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.
- [87] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.
- [88] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth. Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, 86(1):79–122, 2016.
- [89] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.
- [90] M. Coenraad, A. Pellicone, D. J. Ketelhut, M. Cukier, J. Plane, and D. Weintrop. Experiencing cybersecurity one game at a time: A systematic review of cybersecurity digital games. *Simulation & Gaming*, 51(5):586–611, 2020.
- [91] R. Cohen, T. Newton-John, and A. Slater. The relationship between facebook and instagram appearance-focused activities and body image concerns in young women. *Body image*, 23:183–187, 2017.
- [92] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.
- [93] A. Collins, D. Tkaczyk, A. Aizawa, and J. Beel. Position bias in recommender systems for digital libraries. In *International Conference on Information*, pages 335–344. Springer, 2018.
- [94] P. Constantinides, O. Henfridsson, and G. G. Parker. Introduction—platforms and infrastructures in the digital age, 2018.
- [95] A. Corner, L. Whitmarsh, and D. Xenias. Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation. *Climatic change*, 114(3):463–478, 2012.
- [96] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing? how recommender system interfaces affect users’ opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592, 2003.
- [97] R. Costanza, I. Kubiszewski, E. Giovannini, H. Lovins, J. McGlade, K. E. Pickett, K. V. Ragnarsdóttir, D. Roberts, R. De Vogli, and R. Wilkinson. Development: Time to leave gdp behind. *Nature News*, 505(7483):283, 2014.

- [98] M. Costello, J. Hawdon, C. Bernatzky, and K. Mendes. Social group identity and perceptions of online hate*. *Sociological Inquiry*, 89(3):427–452, 2019.
- [99] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18(5):455–496, 2008.
- [100] P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.
- [101] G. Davies, C. Neudecker, M. Ouellet, M. Bouchard, and B. Ducol. Toward a framework understanding of online programs for countering violent extremism. *Journal for Deradicalization*, 1(6):51–86, 2016.
- [102] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.
- [103] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.
- [104] M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. Modeling confirmation bias and polarization. *Scientific reports*, 7(1):1–9, 2017.
- [105] M. A. DeVito, J. Birnholtz, J. T. Hancock, M. French, and S. Liu. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.
- [106] M. A. DeVito, D. Gergle, and J. Birnholtz. "algorithms ruin everything" # riptwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3163–3174, 2017.
- [107] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [108] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
- [109] E. Diener, R. Lusk, D. DeFour, and R. Flax. Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *JPSP*, 39(3):449, 1980.
- [110] E. Diener and M. E. Seligman. Measure for measure: the case for a national well-being index. *Science & Spirit*, 17(2):36–38, 2006.
- [111] D. DiNucci. Design & new media: Fragmented future-web development faces a process of mitosis, mutation, and natural selection. *PRINT-NEW YORK-*, 53:32–35, 1999.
- [112] M. Eirinaki, J. Gao, I. Varlamis, and K. Tserpes. Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78:413 – 418, 2018.
- [113] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168, 2014.

- [114] K. Elghomary and D. Bouzidi. Dynamic peer recommendation system based on trust model for sustainable social tutoring in moocs. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–9. IEEE, 2019.
- [115] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.
- [116] D. Embarcadero-Ruiz, H. Gómez-Adorno, A. Embarcadero-Ruiz, and G. Sierra. Graph-based siamese network for authorship verification. *Mathematics*, 10(2):277, 2022.
- [117] T. Emily, S. Veronica, B. Johanna, S.-R. J. Roberto, S. Lidia, L. Francesco, A. Farbod, O. Dimitri, T. Davide, H.-L. Davinia, and S. Eimler. Empirically investigating virtual learning companions to enhance social media literacy. *Fulantelli et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022.*, 2022.
- [118] L. Engel, A. Chudyk, M. Ashe, H. McKay, D. Whitehurst, and S. Bryan. Older adults’ quality of life—exploring the role of the built environment and social cohesion in community-dwelling seniors on low income. *Social Science & Medicine*, 164:1–11, 2016.
- [119] M. Erdt and C. Rensing. Evaluating recommender algorithms for learning using crowdsourcing. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 513–517. IEEE, 2014.
- [120] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.
- [121] P. Escamilla-Fajardo, M. Alguacil, and S. López-Carril. Incorporating tiktok in higher education: Pedagogical perspectives from a corporal expression sport sciences course. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 28:100302, 2021.
- [122] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig. " i always assumed that i wasn’t really that close to [her]" reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 153–162, 2015.
- [123] D. S. Evans and R. Schmalensee. *Matchmakers: The new economics of multisided platforms*. Harvard Business Review Press, 2016.
- [124] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426, 2019.
- [125] W. Fan, X. Zhao, X. Chen, J. Su, J. Gao, L. Wang, Q. Liu, Y. Wang, H. Xu, L. Chen, et al. A comprehensive survey on trustworthy recommender systems. *arXiv preprint arXiv:2209.10117*, 2022.
- [126] A. Farbod, M. Nils, L. Francesco, D. Gregor, O. Dimitri, K. Udo, H.-L. Davinia, F. Giovanni, and H. H. Ulrich. The “courage companion” - an ai-supported environment for training teenagers in handling social media critically and responsibly. *Fulantelli et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022.*, 2022.
- [127] A. Fedorov. *Media Literacy Education*. ICO: Information for all, 01 2015.

- [128] T. Fillitz and A. J. Saris. *Debating authenticity: Concepts of modernity in anthropological perspective*. Berghahn Books, 2012.
- [129] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017.
- [130] J. W. Fleenor. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies and nations. *Personnel Psychology*, 59(4):982, 2006.
- [131] V. Franchina, M. Vanden Abeele, A. J. Van Rooij, G. Lo Coco, and L. De Marez. Fear of missing out as a predictor of problematic social media use and phubbing behavior among flemish adolescents. *International journal of environmental research and public health*, 15(10):2319, 2018.
- [132] B. L. Fredrickson. The broaden-and-build theory of positive emotions. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1449):1367–1377, 2004.
- [133] N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.
- [134] S. Fu, H. Li, Y. Liu, H. Pirkkalainen, and M. Salo. Social media overload, exhaustion, and use discontinuance: Examining the effects of information overload, system feature overload, and social overload. *Information Processing & Management*, 57(6):102307, 2020.
- [135] N. Fuhr et al. An information nutritional label for online documents. In *ACM SIGIR Forum*, volume 51, pages 46–66. ACM New York, NY, USA, 2018.
- [136] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.
- [137] S. Gai, F. Zhao, Y. Kang, Z. Chen, D. Wang, and A. Tang. Deep transfer collaborative filtering for recommender systems. In *Pacific Rim International Conference on Artificial Intelligence*, pages 515–528. Springer, 2019.
- [138] F. Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.
- [139] K. Garimella, A. Gionis, N. Parotsidis, and N. Tatti. Balancing information exposure in social networks. *arXiv preprint arXiv:1709.01491*, 2017.
- [140] R. K. Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285, 2009.
- [141] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260, 2010.
- [142] J. Gerson. *Social media use and subjective well-being: an investigation of individual differences in personality, social comparison and Facebook behaviour*. PhD thesis, City, University of London, 2018.
- [143] D. Geschke, J. Lorenz, and P. Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1):129–149, 2019.
- [144] A. Gibson. Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media+ Society*, 5(1):2056305119832588, 2019.

- [145] N. K. Gill, R. Garg, and E. A. Doegar. A review paper on digital image forgery detection techniques. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2017.
- [146] N. Gillani, A. Yuan, et al. Me, my echo chamber, and i: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, pages 823–831, 2018.
- [147] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.
- [148] P. Gilster and P. Glistler. *Digital literacy*. Wiley Computer Pub. New York, 1997.
- [149] C. Gimbrone, L. M. Bates, S. J. Prins, and K. M. Keyes. The politics of depression: Diverging trends in internalizing symptoms among us adolescents by political beliefs. *SSM-mental health*, 2:100043, 2022.
- [150] B. Gleason and S. Von Gillern. Digital citizenship with social media: Participatory practices of teaching and learning in secondary education. *Journal of Educational Technology & Society*, 21(1):200–212, 2018.
- [151] D. gnibene, G. Donabauer, U. Kruschwitz, R. S. Wilkens, S. Bursic, D. Hernandez-Leo, E. Theophilou, F. Lomonaco, and U. Kruschwitz. Moving beyond benchmarks and competitions: Towards addressing social media challenges in an educational context. *Datenbank-Spektrum*, 22(1):5–15, 2023.
- [152] P. Golden and D. Danks. Ethical obligations to provide novelty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 502–508, 2021.
- [153] C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.
- [154] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [155] R. Gorwa. What is platform governance? *Information, communication & society*, 22(6):854–871, 2019.
- [156] T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Communications Surveys & Tutorials*, 3(4):2–16, 2000.
- [157] L. Graves and F. Cherubini. *The rise of fact-checking sites in Europe*. Reuters Institute for the Study of Journalism, 2016.
- [158] J. Greer and M. Mark. Evaluation methods for intelligent tutoring systems revisited. *IJAIE*, 26(1):387–392, 2016.
- [159] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is: Evading hate speech detection. In *PWAIS-ACM’18*, pages 2–12. ACM, 2018.
- [160] L. Guarnera, O. Giudice, and S. Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020.
- [161] V. Guliciuc. Complexity and social media. *Procedia-social and behavioral sciences*, 149:371–375, 2014.
- [162] A. Gunawardana, G. Shani, and S. Yogev. Evaluating recommender systems. In *Recommender systems handbook*, pages 547–601. Springer, 2022.

- [163] C. N. Gunawardena. Social presence theory and implications for interaction and collaborative learning in computer conferences. *IJET*, 1(2):147–166, 1995.
- [164] X. Guo, B. Zhu, L. F. Polanía, C. Boncelet, and K. E. Barner. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 635–639, 2018.
- [165] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli. An attention model for group-level emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 611–615, 2018.
- [166] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514, 2013.
- [167] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- [168] W. Han, C. Huang, and J. Yang. Opinion clusters in a modified hegselmann–krause model with heterogeneous bounded confidences and stubbornness. *Physica A: Statistical Mechanics and its Applications*, 531:121791, 2019.
- [169] P. Hartl and U. Kruschwitz. University of regensburg at checkthat! 2021: Exploring text summarization for fake news detection. In *Proceedings of the 12th Conference and Labs of the Evaluation Forum (CLEF2021)*, pages 508–519. CEUR Workshop Proceedings (CEUR-WS.org), 2021.
- [170] P. Hartl and U. Kruschwitz. Applying automatic text summarization for fake news detection. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2702–2713, Marseille, France, June 2022. European Language Resources Association.
- [171] S. Hashida, K. Tamura, and T. Sakai. Classifying tweets using convolutional neural networks with multi-channel distributed representation. *IAENG International Journal of Computer Science*, 46(1):68–75, 2019.
- [172] R. Hegselmann, U. Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.
- [173] I. Heimbach, J. Gottschlich, and O. Hinz. The value of user’s facebook profile data for product recommendation generation. *Electronic Markets*, 25(2):125–138, 2015.
- [174] J. F. Helliwell. How’s life? combining individual and national variables to explain subjective well-being. *Economic modelling*, 20(2):331–360, 2003.
- [175] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020.
- [176] R. Hertwig and T. Grüne-Yanoff. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6):973–986, 2017.
- [177] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- [178] J. Hoffmann and U. Kruschwitz. Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings. In *EVALITA*, 2020.

- [179] D. Holtz, B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral. The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76, 2020.
- [180] S. Horwood and J. Anglim. Personality and problematic smartphone use: A facet-level analysis using the five factor model and hexaco frameworks. *Computers in Human Behavior*, 85:349–359, 2018.
- [181] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.
- [182] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, and D. J. Watts. Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences*, 118(32):e2101967118, 2021.
- [183] A. House. Social media, self-harm and suicide. *BJPsych bulletin*, 44(4):131–133, 2020.
- [184] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening closed regimes: what was the role of social media during the arab spring? *Available at SSRN 2595096*, 2011.
- [185] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee. Deep fake image detection based on pairwise learning. *Applied Sciences*, 10(1):370, 2020.
- [186] D. Hu, X. Zhang, Y. Fan, Z.-Q. Zhao, L. Wang, X. Wu, and X. Wu. On digital image trustworthiness. *Applied Soft Computing*, 48:240–253, 2016.
- [187] M. Huh, A. Liu, A. Owens, and A. A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.
- [188] L. Iaquinta, M. De Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino. Introducing serendipity in a content-based recommender system. In *2008 eighth international conference on hybrid intelligent systems*, pages 168–173. IEEE, 2008.
- [189] A. IBM. The real-world use of big data. *Institute for Business Value, New York, NY*, 2012.
- [190] A. Iovine, P. Lops, F. Narducci, M. de Gemmis, and G. Semeraro. An empirical evaluation of active learning strategies for profile elicitation in a conversational recommender system. *Journal of Intelligent Information Systems*, 58(2):337–362, 2022.
- [191] J. Isaak and M. J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.
- [192] S. Iyengar and S. J. Westwood. Fear and loathing across party lines: New evidence on group polarization. *American journal of political science*, 59(3):690–707, 2015.
- [193] M. Jakesch, H. Leder, and M. Forster. Image ambiguity and fluency. *PLoS One*, 8(9):e74084, 2013.
- [194] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.
- [195] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2016.
- [196] L. M. Jones and K. J. Mitchell. Defining and measuring youth digital citizenship. *New media & society*, 18(9):2063–2079, 2016.

- [197] T. Jung, S. Kim, and K. Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020.
- [198] D. M. Kahan, D. A. Hoffman, D. Braman, and D. Evans. They saw a protest: Cognitive illiberalism and the speech-conduct distinction. *Stan. L. Rev.*, 64:851, 2012.
- [199] M. Kaminskis and D. Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42, 2016.
- [200] M. L. Kapeller, G. Jäger, and M. Füllsack. Homophily in networked agent-based models: a method to generate homophilic attribute distributions to improve upon random distribution approaches. *Computational Social Networks*, 6(1):1–18, 2019.
- [201] V. Karasavva and A. Noorbhai. The real threat of deepfake pornography: A review of canadian policy. *Cyberpsychology, Behavior, and Social Networking*, 24(3):203–209, 2021.
- [202] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. L. Arendt, S. Shaikh, and W. Dou. Vulnerable to misinformation? verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 312–323, 2019.
- [203] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.
- [204] B. Keles, N. McCrae, and A. Grealish. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, 25(1):79–93, 2020.
- [205] Keras. Layer weight initializers. <https://keras.io/api/layers/initializers/>, 2021.
- [206] D. Khattar, J. S. Goud, M. Gupta, and V. Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.
- [207] P. Khosravi, A. Rezvani, and A. Wiewiora. The impact of technology on older adults’ social isolation. *Computers in Human Behavior*, 63:594–603, 2016.
- [208] M. Khwaja, M. Ferrer, J. O. Iglesias, A. A. Faisal, and A. Matic. Aligning daily activities with personality: towards a recommender system for improving wellbeing. In *Proceedings of the 13th acm conference on recommender systems*, pages 368–372, 2019.
- [209] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*, 2016.
- [210] H. H.-s. Kim. The impact of online social networking on adolescent psychological well-being (wb): a population-level analysis of korean school-aged children. *International Journal of Adolescence and Youth*, 22(3):364–376, 2017.
- [211] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics*, 26(1):89–120, 2020.
- [212] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [213] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

- [214] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [215] C. Klein. Social capital or social cohesion: what matters for subjective well-being? *Social Indicators Research*, 110(3):891–911, 2013.
- [216] S. Kopeinik, E. Lex, P. Seitlinger, D. Albert, and T. Ley. Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 409–418, 2017.
- [217] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.
- [218] A. Kozyreva, S. Lewandowsky, and R. Hertwig. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21(3):103–156, 2020.
- [219] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.
- [220] E. Kross, P. Verduyn, E. Demiralp, J. Park, D. S. Lee, N. Lin, H. Shablack, J. Jonides, and O. Ybarra. Facebook use predicts declines in subjective well-being in young adults. *PloS one*, 8(8):e69841, 2013.
- [221] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.
- [222] R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May 2020. European Language Resources Association (ELRA).
- [223] Z. Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.
- [224] D. J. Kuss and M. D. Griffiths. Online social networking and addiction—a review of the psychological literature. *International journal of environmental research and public health*, 8(9):3528–3552, 2011.
- [225] A. Lambrecht and C. E. Tucker. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018)*, 2018.
- [226] R. Latha and R. Nadarajan. Analysing exposure diversity in collaborative recommender systems—entropy fusion approach. *Physica A: Statistical Mechanics and Its Applications*, 533:122052, 2019.
- [227] E. Le Merrer, B. Morgan, and G. Trédan. Setting the record straighter on shadow banning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.
- [228] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. Springer, 2016.
- [229] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang. Fairness in recommendation: A survey. *arXiv preprint arXiv:2205.13619*, 2022.
- [230] Y. Li and Y. Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

- [231] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.
- [232] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- [233] H. Linmei, T. Yang, C. Shi, H. Ji, and X. Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, 2019.
- [234] D. Liu, P. Cheng, H. Zhu, Z. Dong, X. He, W. Pan, and Z. Ming. Mitigating confounding bias in recommendation via information bottleneck. In *Fifteenth ACM Conference on Recommender Systems*, pages 351–360, 2021.
- [235] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [236] E. Llansó, J. Van Hoboken, P. Leerssen, and J. Harambam. Artificial intelligence, content moderation, and freedom of expression. 2020.
- [237] R. Lobo, E. Theophilou, R. Sánchez-Reina, and D. Hernández-Leo. Evaluating an adaptive intervention in collaboration scripts deconstructing body image narratives in a social media educational platform. In *27th International Conference, CollabTech*. Springer, 2022.
- [238] F. Lomonaco, G. Donabauer, M. Siino, et al. Courage at checkthat! 2022: Harmful tweet detection using graph neural networks and electra. *Working Notes of CLEF*, 1, 2022.
- [239] F. Lomonaco, D. Taibi, V. Trianni, and D. Ognibene. A game-based educational experience to increase awareness about the threats of social media filter bubbles and echo chambers inspired by “wisdom of the crowd”: preliminary results. In *Book of Abstracts*, page 84, 2022.
- [240] P. Lops, M. d. Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105, 2011.
- [241] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22):9020–9025, 2011.
- [242] P. Lorenz-Spreen, S. Lewandowsky, C. R. Sunstein, and R. Hertwig. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 2020.
- [243] S. Loughnan, A. Pina, E. A. Vasquez, and E. Puvia. Sexual objectification increases rape victim blame and decreases perceived suffering. *PWQ*, 37(4):455–461, 2013.
- [244] P. B. Lowry, J. Zhang, C. Wang, and M. Siponen. Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, 27(4):962–986, 2016.
- [245] F. Lu, A. Dumitrache, and D. Graus. Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 145–153, 2020.
- [246] C. Mair, A. V. D. Roux, and J. D. Morenoff. Neighborhood stressors and social support as predictors of depressive symptoms in the chicago community adult health study. *Health & place*, 16(5):811–819, 2010.

- [247] N. Manouselis, H. Drachler, R. Vuorikari, H. Hummel, and R. Koper. Recommender systems in technology enhanced learning. In *Recommender systems handbook*, pages 387–415. Springer, 2011.
- [248] M. Mansoury, H. Abdollahpouri, J. Smith, A. Dehpanah, M. Pechenizkiy, and B. Mobasher. Investigating potential factors associated with gender discrimination in collaborative recommender systems. In *The Thirty-Third International Flairs Conference*, 2020.
- [249] E. March and J. Springer. Belief in conspiracy theories: The predictive role of schizotypy, machiavellianism, and primary psychopathy. *PloS one*, 14(12):e0225964, 2019.
- [250] F. S. Marcondes, A. Gala, D. Durães, F. Moreira, J. J. Almeida, V. Baldi, and P. Novais. A profile on twitter shadowban: an ai ethics position paper on free-speech. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 397–405. Springer, 2021.
- [251] D. Marengo, C. Longobardi, M. A. Fabris, and M. Settanni. Highly-visual social media and internalizing symptoms in adolescence: The mediating role of body image concerns. *Computers in Human Behavior*, 82:63–69, 2018.
- [252] I. Marzoli, A. Colantonio, C. Fazio, M. Giliberti, U. S. di Uccio, and I. Testa. Effects of emergency remote instruction during the covid-19 pandemic on university physics students in italy. *Physical Review Physics Education Research*, 17(2):020130, 2021.
- [253] A. Matakos, C. Aslay, E. Galbrun, and A. Gionis. Maximizing the diversity of exposure in a social network. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.
- [254] P. Mavrodiev and F. Schweitzer. The ambiguous role of social influence on the wisdom of crowds: An analytic approach. *Physica A: Statistical Mechanics and its Applications*, 567:125624, 2021.
- [255] M. Mazziotta and A. Pareto. Everything you always wanted to know about normalization (but were afraid to ask). *Rivista Italiana di Economia Demografia e Statistica*, 75(1), 2021.
- [256] S. E. McComb and J. S. Mills. A systematic review on the effects of media disclaimers on young women’s body image and mood. *Body image*, 32:34–52, 2020.
- [257] A. McCosker. Making sense of deepfakes: Socializing ai and building data literacy on github and youtube. *New Media & Society*, page 14614448221093943, 2022.
- [258] S. McGrew and V. L. Byrne. Who is behind this? preparing high school students to evaluate online content. *Journal of Research on Technology in Education*, 53(4):457–475, 2020.
- [259] K. Mehari, A. Farrell, and A.-T. Le. Cyberbullying among adolescents: Measures in search of a construct. *Psychology of Violence*, 4:399–415, 10 2014.
- [260] P. Mena. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & internet*, 12(2):165–183, 2020.
- [261] E. M. Meyers, I. Erickson, and R. V. Small. Digital literacy and informal learning environments: an introduction. *Learning, Media and Technology*, 38(4):355–367, 2013.
- [262] A. C. Michalos. Education, happiness and wellbeing. In *Connecting the quality of life theory to health, well-being and education*, pages 277–299. Springer, 2017.
- [263] S. Milano, M. Taddeo, and L. Floridi. Recommender systems and their ethical challenges. *Ai & Society*, 35(4):957–967, 2020.

- [264] S. Milano, M. Taddeo, and L. Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, 37(1):35–45, 2021.
- [265] S. Milli, L. Belli, and M. Hardt. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 714–722, 2021.
- [266] S. Mills, S. Lucas, L. Irakliotis, M. Rappa, T. Carlson, and B. Perlowitz. Demystifying big data: a practical guide to transforming the business of government. *TechAmerica Foundation, Washington*, 2012.
- [267] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.
- [268] M. Mitchell, J. Lebow, R. Uribe, H. Grathouse, and W. Shoger. Internet use, happiness, social support and introversion: A more fine grained analysis of person variables and internet activity. *Computers in Human Behavior*, 27(5):1857–1861, 2011.
- [269] M. Mladenović, V. Ošmjanski, and S. V. Stanković. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1), Jan. 2021.
- [270] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, and M. Zampieri. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, pages 1–3, 2021.
- [271] J. Mökander and L. Floridi. Operationalising ai governance through ethics-based auditing: an industry case study. *AI and Ethics*, pages 1–18, 2022.
- [272] J. Möller, D. Trilling, N. Helberger, and B. van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.
- [273] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, et al. Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 668–669, 2020.
- [274] K. Muhammad, J. Ahmad, S. Rho, and S. W. Baik. Image steganography for authenticity of visual contents in social networks. *Multimedia Tools and Applications*, 76(18):18985–19004, 2017.
- [275] D. Muise, H. Hosseinmardi, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Quantifying partisan news diets in web and tv audiences. *Science Advances*, 8(28):eabn0083, 2022.
- [276] V. C. Müller. Ethics of artificial intelligence and robotics. In *Stanford Encyclopedia of Philosophy*. Stanford University, 2020.
- [277] J. Nagle. Twitter, cyber-violence, and the need for a critical social media literacy in teacher education: A review of the literature. *Teaching and Teacher Education*, 76:86–94, 2018.
- [278] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, and J. Beltrán. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF ’2022*, Bologna, Italy, 2022.

- [279] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghoulani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulakov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, and J. Köhler. Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, CLEF '2022, Bologna, Italy, 2022.
- [280] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghoulani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulakov, Y. S. Kartal, and J. Beltrán. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørnvåg, and V. Setty, editors, *Advances in Information Retrieval*, pages 416–428, Cham, 2022. Springer International Publishing.
- [281] K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, and Z. Talat, editors. *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.
- [282] K. Narang, Y. Song, A. Schwing, and H. Sundaram. Fuserec: fusing user and item homophily modeling with temporal recommender systems. *Data Mining and Knowledge Discovery*, 35(3):837–862, 2021.
- [283] J. Navajas, T. Niella, G. Garbulsky, B. Bahrami, and M. Sigman. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2):126–132, 2018.
- [284] E. J. Newman, M. Garry, D. M. Bernstein, J. Kantner, and D. S. Lindsay. Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19(5):969–974, 2012.
- [285] A. Y. Ng, S. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.
- [286] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.
- [287] S. J. Nightingale, K. A. Wade, and D. G. Watson. Can people identify original and manipulated photos of real-world scenes? *Cognitive research: principles and implications*, 2(1):1–21, 2017.
- [288] A. N. Nikolakopoulos, X. Ning, C. Desrosiers, and G. Karypis. Trust your neighbors: a comprehensive survey of neighborhood-based methods for recommender systems. *Recommender Systems Handbook*, pages 39–89, 2022.
- [289] N. J. Nilsson and N. J. Nilsson. *Artificial intelligence: a new synthesis*. Morgan Kaufmann, 1998.
- [290] H. Noorazar, K. R. Vixie, A. Talebanpour, and Y. Hu. From classical to modern opinion dynamics. *International Journal of Modern Physics C*, 31(07):2050101, 2020.
- [291] C. Novelli, M. Taddeo, and L. Floridi. *Accountability in artificial intelligence: what it is and how it works*. Springer, 2022.
- [292] S. Nuere and L. De Miguel. The digital/technological connection with covid-19: An unprecedented challenge in university teaching. *Technology, Knowledge and Learning*, 26(4):931–943, 2021.
- [293] D. S. Nunes, P. Zhang, and J. S. Silva. A survey on human-in-the-loop applications towards an internet of all. *IEEE Communications Surveys & Tutorials*, 17(2):944–965, 2015.

- [294] J. A. Obar and S. S. Wildman. Social media definition and the governance challenge—an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy*, 39(9):745–750, 2015.
- [295] D. Ognibene, V. G. Fiore, and X. Gu. Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. *Neural Networks*, 116:269–278, 2019.
- [296] D. Ognibene, D. Taibi, U. Kruschwitz, R. S. Wilkens, D. Hernandez-Leo, E. Theophilou, L. Scifo, R. A. Lobo, F. Lomonaco, S. Eimler, et al. Challenging social media threats using collective well-being aware recommendation algorithms and an educational virtual companion. *arXiv preprint arXiv:2102.04211*, 2021.
- [297] D. Ognibene, R. Wilkens, D. Taibi, D. Hernández-Leo, U. Kruschwitz, G. Donabauer, E. Theophilou, F. Lomonaco, S. Bursic, R. A. Lobo, and et al. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence*, Dec 2022.
- [298] K. O’Hara and D. Stevens. Echo chambers and online radicalism: Assessing the internet’s complicity in violent extremism. *Policy & Internet*, 7(4):401–422, 2015.
- [299] W. H. Organization et al. Infodemic management: an overview of infodemic management during covid-19, january 2020–may 2021, 2021.
- [300] L. Osberg. On the limitations of some current usages of the gini index. *Review of Income and Wealth*, 63(3):574–584, 2017.
- [301] E. Ostrom. The difference: How the power of diversity creates better groups, firms, schools, and societies. by scott e. page. princeton: Princeton university press, 2007. 448p. 19.95 paper. *Perspectives on Politics*, 6(4):828–829, 2008.
- [302] J. F. Padgett and C. K. Ansell. Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, 98(6):1259–1319, 1993.
- [303] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- [304] A. Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1990.
- [305] Y. Pan, F. He, and H. Yu. Learning social representations with deep autoencoder for recommender system. *World Wide Web*, 23(4):2259–2279, 2020.
- [306] A. Panciroli. *2k10 L’anno zero dei social*. Sperling & Kupfer, 2021.
- [307] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 885–894, 2020.
- [308] E. Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.
- [309] H. Parlak Sert and H. Başkale. Students’ increased time spent on social media, and their level of coronavirus anxiety during the pandemic predict increased social media addiction. *Health Information & Libraries Journal*, 2022.
- [310] J. Pavlopoulos, L. Laugier, J. Sorensen, and I. Androutopoulos. Semeval-2021 task 5: Toxic spans detection. *Proceedings of SemEval*, 2021.

- [311] S. T. Peddinti, K. W. Ross, and J. Cappos. "on the internet, nobody knows you're a dog" a twitter case study of anonymity in social networks. In *Proceedings of the second ACM conference on Online social networks*, pages 83–94, 2014.
- [312] G. Pennycook, J. A. Cheyne, N. Barr, D. J. Koehler, and J. A. Fugelsang. Bullshit receptivity scale. *Judgment and Decision Making*, 2015.
- [313] G. Pennycook and D. G. Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200, 2020.
- [314] N. Perra and L. E. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific reports*, 9(1):1–11, 2019.
- [315] M. Peschl, A. Zgonnikov, F. A. Oliehoek, and L. C. Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. *arXiv preprint arXiv:2201.00012*, 2021.
- [316] S. Plous. *The psychology of judgment and decision making*. McGraw-Hill Book Company, 1993.
- [317] M. Polignano, C. Musto, M. de Gemmis, P. Lops, and G. Semeraro. Together is better: Hybrid recommendations combining graph embeddings and contextualized word representations. In *Fifteenth ACM Conference on Recommender Systems*, pages 187–198, 2021.
- [318] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.
- [319] T. Postmes and R. Spears. *Deindividuation and antinormative behavior: A meta-analysis.*, volume 123. American Psychological Association, 1998.
- [320] W. J. Potter. The state of media literacy. *Journal of broadcasting & electronic media*, 54(4):675–696, 2010.
- [321] P. Pradhyumna, G. Shreya, et al. Graph neural network (gnn) in image and video understanding using deep learning for computer vision applications. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1183–1189. IEEE, 2021.
- [322] C. S. Prakash, A. Kumar, S. Maheshkar, and V. Maheshkar. An integrated method of copy-move and splicing for image forgery detection. *Multimedia Tools and Applications*, 77(20):26939–26963, 2018.
- [323] M. Prior. Visual political knowledge: A different road to competence? *The Journal of Politics*, 76(1):41–57, 2014.
- [324] A. K. Przybylski, K. Murayama, C. R. DeHaan, and V. Gladwell. Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in human behavior*, 29(4):1841–1848, 2013.
- [325] A. K. Purohit, L. Barclay, and A. Holzer. Designing for digital detox: Making social media less addictive with digital nudges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.
- [326] X. Qiu, D. FM Oliveira, A. Sahami Shirazi, A. Flammini, and F. Menczer. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):1–7, 2017.
- [327] W. Quattrociocchi, G. Caldarelli, and A. Scala. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports*, 4(1):1–7, 2014.
- [328] W. Quattrociocchi, A. Scala, and C. R. Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

- [329] K. Rajkumar, G. Saint-Jacques, I. Bojinov, E. Brynjolfsson, and S. Aral. A causal test of the strength of weak ties. *Science*, 377(6612):1304–1310, 2022.
- [330] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [331] N. Ranjbar Kermany, W. Zhao, J. Yang, J. Wu, and L. Pizzato. A fairness-aware multi-stakeholder recommender system. *World Wide Web*, 24(6):1995–2018, 2021.
- [332] B. Rastegarpanah, K. P. Gummadi, and M. Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 231–239, 2019.
- [333] R. Ray. Prediction markets and the financial" wisdom of crowds". *The Journal of Behavioral Finance*, 7(1):2–4, 2006.
- [334] E. Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.
- [335] M. Ribble. *Digital citizenship in schools: Nine elements all students should know*. International Society for Technology in Education, 2015.
- [336] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [337] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf. Quantifying information overload in social media and its impact on social contagions. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- [338] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [339] C. Romero and S. Ventura. Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(1):e1187, 2017.
- [340] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.
- [341] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer. Assessing social presence in asynchronous text-based computer conferencing. *The Journal of Distance Education/Revue de l'education Distance*, 14(2):50–71, 1999.
- [342] B. Roy, C. Riley, L. Sears, and E. Y. Rula. Collective well-being to improve population health outcomes: an actionable conceptual model and review of the literature. *American Journal of Health Promotion*, 32(8):1800–1813, 2018.
- [343] B. Rozemberczki and R. Sarkar. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, page 1325–1334. ACM, 2020.
- [344] B. Rozemberczki and R. Sarkar. Twitch gamers: a dataset for evaluating proximity preserving and structural role-based node embeddings, 2021.
- [345] V. Rubin and T. Lukoianova. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1):4, 2013.

- [346] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [347] L. Ruiz, F. Gama, and A. Ribeiro. Gated graph recurrent neural networks. *IEEE Transactions on Signal Processing*, 68:6303–6318, 2020.
- [348] T. Ryan, K. A. Allen, D. L. Gray, and D. M. McInerney. How social are social media? a review of online social behaviour and connectedness. *Journal of Relationships Research*, 8, 2017.
- [349] S. Sadeghi, S. Dadkhah, H. A. Jalab, G. Mazzola, and D. Uliyan. State of the art in passive digital image forgery detection: copy-move image forgery. *Pattern Analysis and Applications*, 21(2):291–306, 2018.
- [350] A. L. Sánchez Laws and T. Utne. Ethics guidelines for immersive journalism. *Frontiers in Robotics and AI*, 6:28, 2019.
- [351] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22:4349–4357, 2014.
- [352] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, and I. Pisa. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA*, 2020.
- [353] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402, 2021.
- [354] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.
- [355] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho. Humans are easily fooled by digital images. *Computers & Graphics*, 68:142–151, 2017.
- [356] A. Schlesinger, E. Chandrasekharan, C. A. Masden, A. S. Bruckman, W. K. Edwards, and R. E. Grinter. Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 6912–6924, 2017.
- [357] T. Schulenkorf, V. Krah, K. Dadaczynski, and O. Okan. Addressing health literacy in schools in germany: concept analysis of the mandatory digital and media literacy school curriculum. *Frontiers in Public Health*, 9, 2021.
- [358] C. Schwind, J. Buder, U. Cress, and F. W. Hesse. Preference-inconsistent recommendations: An effective approach for reducing confirmation bias and stimulating divergent thinking? *Computers & Education*, 58(2):787–796, 2012.
- [359] C. A. Scolari, M.-J. Masanet, M. Guerrero-Pico, and M.-J. Establés. Transmedia literacy in the new media ecology: Teens’ transmedia skills and informal learning strategies. *EPI*, 27(4):801–812, 2018.
- [360] M. E. Seligman. *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2012.
- [361] S. Seufert, C. Meier, M. Soellner, and R. Rietsche. A pedagogical perspective on big data and learning analytics: A conceptual model for digital learning support. *Technology, Knowledge and Learning*, 24(4):599–619, 2019.

- [362] Y. Shao and Y.-h. Xie. Research on cold-start problem of collaborative filtering algorithm. In *Proceedings of the 2019 3rd International Conference on Big Data Research*, pages 67–71, 2019.
- [363] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O’Brien. Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society*, 21(2):438–463, 2019.
- [364] M. Siino, E. Di Nuovo, T. Ilenia, and M. La Cascia. Detection of hate speech spreaders using convolutional neural networks. In *PAN 2021 Profiling Hate Speech Spreaders on Twitter@CLEF*, volume 2936, pages 2126–2136. CEUR, 2021.
- [365] M. Siino, M. La Cascia, and I. Tinnirello. Whosnext: Recommending twitter users to follow using a spreading activation network based approach. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 62–70. IEEE, 2020.
- [366] M. Siino, M. La Cascia, and I. Tinnirello. McRock at SemEval-2022 Task 4: Patronizing and Condescending Language Detection using Multi-Channel CNN and DistilBERT. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.
- [367] H. A. Simon et al. Designing organizations for an information-rich world. *Computers, communications, and the public interest*, 72:37, 1971.
- [368] A. Sirbu, D. Pedreschi, F. Giannotti, and J. Kertész. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one*, 14(3):e0213246, 2019.
- [369] J. J. Smith and L. Beattie. Recsys fairness metrics: Many to use but which one to choose? *arXiv preprint arXiv:2209.04011*, 2022.
- [370] E. M. Sosu. The development and psychometric validation of a critical thinking disposition scale. *Thinking skills and creativity*, 9:107–119, 2013.
- [371] J. Spratt. Conceptualising wellbeing. In *Wellbeing, Equity and Education*, pages 35–56. Springer, 2017.
- [372] J. Spratt. Wellbeing, equity and education. *A critical analysis of policy discourses of wellbeing in schools*. Cham: Springer, 2017.
- [373] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [374] H. Steck, L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, and J. Basilico. Deep learning for recommender systems: A netflix case study. *AI Magazine*, 42(3):7–18, 2021.
- [375] R. G. Steele, J. A. Hall, and J. L. Christofferson. Conceptualizing digital stress in adolescents and young adults: Toward the development of an empirically based model. *Clinical Child and Family Psychology Review*, 23(1):15–26, 2020.
- [376] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.
- [377] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin. Information gerrymandering and undemocratic decisions. *Nature*, 573(7772):117–121, 2019.

- [378] A. Stoica, C. Riederer, et al. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pages 923–932, 2018.
- [379] C. Stokel-Walker. Covid-19: The countries that have mandatory vaccination for health workers. *bmj*, 373, 2021.
- [380] J. Stray. Designing recommender systems to depolarize. *arXiv preprint arXiv:2107.04953*, 2021.
- [381] J. Stray, A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, L. Beattie, M. Ekstrand, C. Leibowicz, C. M. Sehat, et al. Building human values into recommender systems: An interdisciplinary synthesis. *arXiv preprint arXiv:2207.10192*, 2022.
- [382] F. Strub, R. Gaudel, and J. Mary. Hybrid recommender system based on autoencoders. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 11–16, 2016.
- [383] J. Su, A. Sharma, and S. Goel. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*, pages 1157–1167, 2016.
- [384] W. Su, G. Chen, and Y. Hong. Noise leads to quasi-consensus of hegselmann–krause opinion dynamics. *Automatica*, 85:448–454, 2017.
- [385] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022.
- [386] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021.
- [387] Z. Sun, I. Lorscheid, J. D. Millington, S. Lauf, N. R. Magliocca, J. Groeneveld, S. Balbi, H. Nolzen, B. Müller, J. Schulze, et al. Simple or complicated agent-based models? a complicated issue. *Environmental Modelling & Software*, 86:56–67, 2016.
- [388] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [389] B. Swire-Thompson, J. DeGutis, and D. Lazer. Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition*, 9(3):286–299, 2020.
- [390] D. M. Szymanski, L. B. Moffitt, and E. R. Carr. Sexual objectification of women: Advances to theory and research. *The Counseling Psychologist*, 39(1):6–38, 2011.
- [391] J. Sánchez-Reina, D. Hernández-Leo, E. Theophilou, and R. Lobo-Quintero. But i don’t wanna share my data’. analyzing teen’s concerns about the use of social media. In *IAMCR 2022 Conference. Communication Research in the Era of Neo-Globalisation: Reorientations, Challenges and Changing Contexts (Beijing and Online)*, 2022.
- [392] F. Tahmasebi, M. Meghdadi, S. Ahmadian, and K. Valiollahi. A hybrid recommendation system based on profile expansion technique to alleviate cold start problem. *Multimedia Tools and Applications*, 80(2):2339–2354, 2021.
- [393] D. Taibi, G. Fulantelli, V. Monteleone, D. Schicchi, and L. Scifo. An innovative platform to promote social media literacy in school contexts. In *ECEL 2021 20th European Conference on e-Learning*, page 460. Academic Conferences International limited, 2021.

- [394] J. K. Tarus, Z. Niu, and A. Yousif. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generation Computer Systems*, 72:37–48, 2017.
- [395] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin, 2009.
- [396] A. Thierer. The perils of classifying social media platforms as public utilities. *CommLaw Conspectus*, 21:249, 2012.
- [397] P. B. Thorat, R. M. Goudar, and S. Barve. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4):31–36, 2015.
- [398] M. Tomlein, B. Pecher, J. Simko, I. Srba, R. Moro, E. Stefancova, M. Kompan, A. Hrcakova, J. Podrouzek, and M. Bielikova. An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes. In *Fifteenth ACM Conference on Recommender Systems*, pages 1–11, 2021.
- [399] S. Tong. *Active learning: theory and applications*. Stanford University, 2001.
- [400] C. W. Topp, S. D. Østergaard, S. Søndergaard, and P. Bech. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and psychosomatics*, 84(3):167–176, 2015.
- [401] P. Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119, 2022.
- [402] H. N. Tran and U. Kruschwitz. ur-iw-hnt at germeval 2021: An ensembling strategy with multiple bert models. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 83–87, Duesseldorf, Germany, Sept. 2021. Association for Computational Linguistics.
- [403] H. N. Tran and U. Kruschwitz. ur-iw-hnt at checkthat! 2022: Cross-lingual text summarization for fake news detection. In *Proceedings of the 13th Conference and Labs of the Evaluation Forum (CLEF2022)*. CEUR Workshop Proceedings (CEUR-WS.org), 2022.
- [404] S. Tseng and B. Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999.
- [405] A. K. Tsitsika, E. C. Tzavela, M. Janikian, K. Ólafsson, A. Iordache, T. M. Schoenmakers, C. Tzavara, and C. Richardson. Online social networking in adolescence: Patterns of use in six european countries and links with psychosocial functioning. *Journal of adolescent health*, 55(1):141–147, 2014.
- [406] C. Turban and U. Kruschwitz. Tackling irony detection using ensemble classifiers and data augmentation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6976–6984, Marseille, France, June 2022. European Language Resources Association.
- [407] C. Vaccari and A. Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1):2056305120903408, 2020.
- [408] T. Valtonen, M. Tedre, K. Mäkitalo, and H. Vartiainen. Media literacy education in the age of machine learning. *Journal of Media Literacy Education*, 11(2):20–36, 2019.
- [409] J.-P. Van Staaldouin and S. de Freitas. A game-based learning framework: Linking game design and learning. *Learning to play*, 53:29, 2011.

- [410] H. P. Vanchinathan, I. Nikolic, F. De Bona, and A. Krause. Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 225–232, 2014.
- [411] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.
- [412] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [413] B. Vecchione, K. Levy, and S. Barocas. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO ’21, New York, NY, USA, 2021. Association for Computing Machinery.
- [414] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [415] P. Verduyn, O. Ybarra, M. Résibois, J. Jonides, and E. Kross. Do social network sites enhance or undermine subjective well-being? a critical review. *SIPR*, 11(1):274–302, 2017.
- [416] V. Verrastro, F. Liga, F. Cuzzocrea, M. C. Gugliandolo, et al. Fear the instagram: beauty stereotypes, body image and instagram use in a sample of male and female adolescents. *Querty-Open and Interdisciplinary Journal of Technology, Culture and Education*, 15(1):31–49, 2020.
- [417] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.
- [418] S. Vrijenhoek, M. Kaya, N. Metoui, J. Möller, D. Odijk, and N. Helberger. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 173–183, 2021.
- [419] A. Walker and L. J. Van Der Maesen. *Social quality: From theory to indicators*. Springer, 2011.
- [420] K. L. Walker. Surrendering information through the looking glass: Transparency, trust, and protection. *Journal of Public Policy & Marketing*, 35(1):144–158, 2016.
- [421] J.-L. Wang, L. A. Jackson, J. Gaskin, and H.-Z. Wang. The effects of social networking site (sns) use on college students’ friendship and well-being. *Computers in Human Behavior*, 37:229–236, 2014.
- [422] R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang. A survey on opinion mining: From stance to product aspect. *IEEE Access*, 7:41101–41124, 2019.
- [423] X. Wang, A. D. Sirianni, S. Tang, Z. Zheng, and F. Fu. Public discourse and social network echo chambers driven by socio-cognitive biases. *Physical Review X*, 10(4):041042, 2020.
- [424] Y. Wang, L. Tao, and X. X. Zhang. Recommending for a multi-sided marketplace with heterogeneous contents. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 456–459, 2022.
- [425] H. Warne, L. Dencik, and A. Hintz. Advancing civic participation in algorithmic decision-making: a guidebook for the public sector, 2021.
- [426] H. Webb, P. Burnap, R. Procter, O. Rana, B. C. Stahl, M. Williams, W. Housley, A. Edwards, and M. Jirotko. Digital wildfires: propagation, verification, regulation, and responsible innovation. *ACM Transactions on Information Systems (TOIS)*, 34(3):15, 2016.

- [427] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2(1):1–9, 2012.
- [428] M. Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019.
- [429] J. White. Wellbeing and education: Issues of culture and authority. *Journal of Philosophy of Education*, 41(1):17–28, 2007.
- [430] E. Whittaker and R. M. Kowalski. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29, 2015.
- [431] R. Wilkens and D. Ognibene. bicourage: ngram and syntax gens for hate speech detection. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*, 2021.
- [432] R. Wilkens and D. Ognibene. Mb-courage@ exist: Gen classification for sexism identification in social networks. *IberLEF@ EXIST*, 2021.
- [433] H. L. Willis and L. Philipson. *Understanding electric utilities and de-regulation*. CRC Press, 2018.
- [434] H. C. Woods and H. Scott. # sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of adolescence*, 51:41–49, 2016.
- [435] H. Wu, Y. Liu, and J. Wang. Review of text classification methods on deep learning. *CMC-Computers, Materials & Continua*, 63(3):1309–1321, 2020.
- [436] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.
- [437] Y. Wu, W. AbdAlmageed, and P. Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019.
- [438] Y.-H. Wu and S.-D. Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [439] X. Xie, Y. Du, and Q. Bai. Why do people resist algorithms? from the perspective of short video usage motivations. *Frontiers in Psychology*, 13, 2022.
- [440] S. Xu, H. H. Yang, J. MacLeod, and S. Zhu. Social media competence and digital citizenship among college students. *Convergence*, 25(4):735–752, 2019.
- [441] S. Yakhchi, A. Beheshti, S.-M. Ghafari, M. A. Orgun, and G. Liu. Towards a deep attention-based sequential recommender system. *IEEE Access*, 8:178073–178084, 2020.
- [442] Y. Yamamoto. A morality based on trust: Some reflections on japanese morality. *Philosophy East and West*, 40(4):451–469, 1990.
- [443] L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.
- [444] S. Yao, Y. Halpern, N. Thain, X. Wang, K. Lee, F. Prost, E. H. Chi, J. Chen, and A. Beutel. Measuring recommender system effects with simulated users. *arXiv preprint arXiv:2101.04526*, 2021.
- [445] A. L. Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

- [446] M. Yesilada and S. Lewandowsky. A systematic review: The youtube recommender system and pathways to problematic content, 2021.
- [447] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.
- [448] F. M. Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.
- [449] J. Zarocostas. How to fight an infodemic. *The lancet*, 395(10225):676, 2020.
- [450] S. Zhang and K. Balog. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, pages 1512–1520, 2020.
- [451] Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [452] Z. Zhao. Analysis on the “douyin (tiktok) mania” phenomenon based on recommendation algorithms. In *E3S Web of Conferences*, volume 235, page 03029. EDP Sciences, 2021.
- [453] L. Zheng, Y. Zhang, and V. L. Thing. A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58:380–399, 2019.
- [454] J. Zhou. Boomerangs versus javelins: how polarization constrains communication on climate change. *Environmental Politics*, 25(5):788–811, 2016.
- [455] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.
- [456] Q. Zhou and Z. Wu. Multidimensional friedkin-johnsen model with increasing stubbornness in social networks. *Information Sciences*, 600:170–188, 2022.
- [457] S. Zimmerman, A. Thorpe, J. Chamberlain, and U. Kruschwitz. Towards search strategies for better privacy and information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20*, pages 124–134. Association for Computing Machinery, 2020.

Acknowledgements

I want to thank the following people who contributed to this research and, more importantly, helped make me who I am.

- Many thanks to **Professor Dimitri Ognibene**, who allowed me to join his lab and patiently supervised me during this hard path and interesting times.
- I would also like to thank my mentor, Professor **Roberto Golinelli**, who inoculated in me the passion for econometrics and research and unpredictably kick-started this amazing journey of the PhD.
- Thanks to **Paolo Marocco**, for sure the best *weak link* of my life, for pushing me to persist during the dark times, whom nowadays I consider a valuable friend.
- Thanks to **Gregor Donabauer and Sathya Bursic**, the best labmates a researcher could ever dream of.
- Thanks to my coauthor **Marco Siino** for sharing his passion and expertise with me. From a professional and human point of view, his advice has been crucial to concluding my PhD.
- Ringrazio mio padre e mia madre per avermi sempre sostenuto e amato, dedico questo obiettivo raggiunto a voi.
- I dedicate this work to the beloved memory of my grandparents (Maria, Adriana, Ezio, Alvaro) and my aunt Giuliana, I'm sure you are celebrating somewhere.
- A big hug to my friends **Edoardo, Luca, MatteoX2, Camilla, Lorenzo, Tommaso, Carolina, Enrica, Valerio, Antonio, Gianmarco** for supporting and enduring me, I know, I am not an easy person!
- Thanks to Alessandra, *You are my heart. Could I live without my heart?*
- Thanks to **Termine-Serpieri family** for always being there and, in particular, to M. for telling me that I would find strength and courage to move beyond any obstacle.
- To my cats, **Violetta, Roberto, Cesare, and Lisa**

The journey of completing this PhD has been a challenging but immensely satisfying one. The process of researching and writing this thesis has been a steep learning curve, but the sense of accomplishment upon its completion is indescribable.

Francesco Lomonaco, Rome, April 7, 2023
