Datenbank-Spektrum

Moving Beyond Benchmarks and Competitions: Towards Addressing Social Media Challenges in an Educational Context

--Manuscript Draft--

Manuscript Number:	DASP-D-22-00023R1
Full Title:	Moving Beyond Benchmarks and Competitions: Towards Addressing Social Media Challenges in an Educational Context
Article Type:	Schwerpunktbeitrag
Corresponding Author:	dimitri ognibene, Ph.D. Università degli Studi di Milano-Bicocca Milano, Lombardia ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Università degli Studi di Milano-Bicocca
Corresponding Author's Secondary Institution:	
First Author:	Dimitri Ognibene, Ph.D.
First Author Secondary Information:	
Order of Authors:	Dimitri Ognibene, Ph.D.
	Gregor Donabauer
	Emily Theophilou
	Sathya Bursic
	Francesco Lomonaco
	Rodrigo Wilkens
	Davinia Hernandez-Leo
	Udo Kruschwitz
Order of Authors Secondary Information:	

Funding Information:		
	Volkswagen Foundation (9B145)	Prof. Dr. Dimitri Ognibene
	Volkswagen Foundation (95564)	Prof. Dr. Udo Kruschwitz
	Volkswagen Foundation (95566)	Prof. Dr. Davinia Hernandez-Leo
	Agencia Estatal de Investigación (PID2020-112584RB-C33/MICIN/AEI/ 10.13039/501100011033)	Prof. Dr. Davinia Hernandez-Leo
	Agencia Estatal de Investigación (MDM-2015-0502)	Prof. Dr. Davinia Hernandez-Leo
	Institució Catalana de Recerca i Estudis Avançats	Prof. Dr. Davinia Hernandez-Leo

Abstract:	Natural language processing and other areas of artificial intelligence have seen staggering progress in recent years, yet much of this is reported with reference to somewhat limited benchmark datasets. We see the deployment of these techniques in realistic use cases as the next step in this development. In particular, much progress is still needed in educational settings, which can strongly improve users' safety on social media. We present our efforts to develop multi-modal machine learning algorithms to be integrated into a social media companion aimed at supporting and educating users in dealing with fake news and other social media threats.

	Inside the companion environment, such algorithms can automatically assess and enable users to contextualize different aspects of their social media experience. They can estimate and display different characteristics of content in supported users' feeds, such as `fakeness' and `sentiment', and suggest related alternatives to enrich users' perspectives. In addition, they can evaluate the opinions, attitudes, and neighbourhoods of the users and of those appearing in their feeds. The aim of the latter process is to raise users' awareness and resilience to filter bubbles and echo chambers, which are almost unnoticeable and rarely understood phenomena that may affect users' information intake unconsciously and are unexpectedly widespread. Social media environment is rapidly changing and complex. While our algorithms show state-of-the-art performance, they rely on task-specific datasets, and their reliability may decrease over time and be limited against novel threats. The negative impact of these limits may be exasperated by users' over-reliance on algorithmic tools. Therefore, companion algorithms and educational activities are meant to increase users' awareness of social media threats while exposing the limits of such algorithms. This will also provide an educational example of the limits affecting the machine-learning components of social media platforms. We aim to devise, implement and test the impact of the companion and connected educational activities in acquiring and supporting conscientious and autonomous social media usage.
Response to Reviewers:	We thank all reviewers for their insightful comments that helped us to improve our work. We tried to adopt all suggested ideas into our manuscript as good as possible.

We thank all reviewers for their insightful comments that helped us to improve our work. We tried to adopt all suggested ideas into our manuscript as good as possible.

Moving Beyond Benchmarks and Competitions: Towards Addressing Social Media Challenges in an Educational Context

Dimitri Ognibene[1*], Gregor Donabauer[1,2], Emily Theophilou[3], Sathya Bursic[1], Francesco Lomonaco[1], Rodrigo Wilkens[4], Davinia Hernández-Leo[3] and Udo Kruschwitz[2]

[1]Dipartimento di Psicologia, Università Milano-Bicocca, Milan, Italy.
[2]Information Science, University of Regensburg, Regensburg, Germany.
[3]Dept. of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain.
[4]Institut Langage et Communication, Université catholique de Louvain, Louvain, Belgium.

*Corresponding author(s). E-mail(s): dimitri.ognibene@unimib.it;
Contributing authors: gregor.donabauer@ur.de; emily.theophilou@upf.edu; sathya.bursic@unimib.it; f.lomonaco5@campus.unimib.it; rodrigo.wilkens@uclouvain.be; davinia.hernandez-leo@upf.edu; udo.kruschwitz@ur.de;

Abstract

Natural language processing and other areas of artificial intelligence have seen staggering progress in recent years, yet much of this is reported with reference to somewhat limited benchmark datasets. We see the deployment of these techniques in realistic use cases as the next step in this development. In particular, much progress is still needed in educational settings, which can strongly improve users' safety on social media. We present our efforts to develop multi-modal machine learning algorithms to be integrated into a social media companion aimed at supporting and educating users in dealing with fake news and other social media threats. Inside the companion environment, such algorithms can automatically assess and enable users to contextualize different aspects of their social media experience. They can estimate and display different characteristics of content in supported users' feeds, such as 'fakeness' and 'sentiment', and suggest related alternatives to enrich users' perspectives. In addition, they can evaluate the opinions, attitudes, and neighbourhoods of the users and of those appearing in their feeds. The aim of the latter process is to raise users' awareness and resilience to filter bubbles and echo chambers, which are almost unnoticeable and rarely understood phenomena that may affect users' information intake unconsciously and are unexpectedly widespread. Social media environment is rapidly changing and complex. While our algorithms show state-of-the-art performance, they rely on task-specific datasets, and their reliability may decrease over time and be limited against novel threats. The negative impact of these limits may be exasperated by users' over-reliance on algorithmic tools. Therefore, companion algorithms and educational activities are meant to increase users' awareness of social media threats while exposing the limits of such algorithms. This will also provide an educational example of the limits affecting the machine-learning components of social media platforms. We aim to devise, implement and test the impact of the companion and connected educational activities in acquiring and supporting conscientious and autonomous social media usage.

Keywords: Social media, Fake news, Hate speech, Toxic content, Education, Companion

1 Introduction

Social media have become an integral part of society in recent years. Besides all the benefits this has brought, it has also uncovered a number of serious problems including the increasing speed and the number of interactions that go beyond the users' ability to monitor and understand such content, resulting in threats such as the pervasive diffusion of fake news and biased as well as toxic content such as hate speech. A common way to address such challenges is through the adoption of natural language processing powerful state-of-the-art approaches, triggered by the paradigmatic shift that the introduction of transformer-based models (such as BERT) has led to (Devlin et al, 2019). The adoption of common benchmark collections has been another major driver in this context. Several of those datasets focus on the detection of single threats (e.g. in the domain of fake news detection (Shu et al, 2020) or for hate speech detection (Mathew et al, 2021)). Others try to unify existing text data collections, e.g. for classification of toxic content (Risch et al, 2021a; Vidgen and Derczynski, 2021).

Such benchmarks have also increasingly been utilized in a growing number of shared tasks and competitions (with leaderboards), primarily led by the machine learning (ML) community. However, a lot of work in this area remains in a purely academic classification scenario and is not being put to use in a practical context. Perhaps more importantly, it has been observed that the performance levels reported for common benchmarks do not necessarily reflect how well the algorithms will work in a realistic use case as systems are often very brittle and the performance levels do not actually transfer easily to different domains, datasets or even variations of the same dataset (Bowman and Dahl, 2021).

Instead of adopting a well-controlled setting (without any real user involvement) we aim to address an actual practical use case (which does not lend itself to being modelled around existing benchmark collections). Our starting point is the observation that social media users often have a limited understanding of the platforms and their algorithms and, more importantly, the effects of their actions on others' experiences and their role in the proliferation of toxic phenomena (Valtonen et al, 2019; Kozyreva et al, 2020). We present a framework that serves as a machine-learning-based social media education tool that aims at integrating solutions to the above-mentioned problems directly in the users' social media experience (Ognibene et al, 2023)[1]. As such the user's feed is augmented automatically with additional information on the content and underlying producing social network, as can be seen in Figure 2. Machine learning is used to trigger personalized and contextualized educational experiences that rise users' awareness about social media and its threats. At the same time, autonomous evaluation is encouraged by highlighting the principles and limits of the involved algorithmic components. The ultimate objective is to educate and empower social media users. Figure 1 gives a high-level view of the educational framework we are proposing.

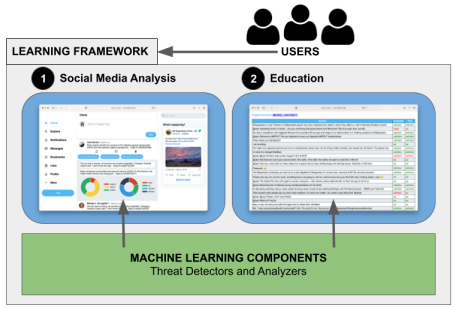

Fig. 1 Conceptual view of our proposed framework. *Social Media Analysis* shows the tool that provides additional information while browsing the feed; *Education* represents educational activities (example here: machine learning models' limitations, described in more detail in section 5.3.

In this paper, we start by discussing threats arising through social media, then present trends in how the community works on solving such issues, and then contextualize these developments in a scenario of practical use taken from the COURAGE project.

2 Social Media Threats

Threats occurring on social media cover a broad range of categories due to the vast amounts of multifaceted content on such platforms. As a result, crucial ethical and practical issues, like preserving

[1]This work is part of the COURAGE project, introducing solutions to social media harm education for teenagers (https://www.upf.edu/web/courage).

Fig. 2 Screenshot of social media content analysis results inside our Twitter demo interface. Here other user posts of the person connected to the first tweet as well as sentiment and emotion analysis are displayed. The buttons under each post allow to show/hide these additional information.

freedom of speech and allowing users to be collectively satisfied while dealing with the conflicts generated by their different opinions and contrasting interests, lead to negative influences on users and society.

Critical cases include the spread of fake news, biased content and the growing trend of hate practices (which indeed is not a new phenomenon on the internet (Gerstenfeld et al, 2003; Schafer, 2002; Chan et al, 2016)). Even though social media platforms are presenting policies against hate speech, discrimination or violent and racist content, the mentioned threats are still part of these websites[2] (Hale, 2012; Bliuc et al, 2018), underlining the need for raising awareness to the users.

Before presenting ways of how to counteract these issues in general, and how we do that with the help of our approach in the COURAGE project, we want to give a brief overview about the categories of social media threats, grouping them in **(1)** content-based, **(2)** algorithmic, **(3)** dynamics, and **(4)** cognitive and socio-emotional.

The transitions between these types of threats are fluid, making it hard to provide clear distinctions. Our focus while describing these issues lies on teenagers, which for example are heavily affected by bullying (Talwar et al, 2014; Mladenović et al, 2021), addiction (Tariq et al, 2012; Shensa et al, 2017), body stereotypes, and others (Mcandrew and Jeong, 2012; Clarke, 2009; Ozimek et al, 2017). This is also the reason why we aim at supporting this exceptionally vulnerable group of the society in the COURAGE project.

2.1 Content-Based Threats

Content-based threats are very common for all types of media, including classical outlets, but

[2]Simon Wiesenthal Center: http://www.digitalhate.net, Online Hate and Harassment Report: The American Experience 2020: https://www.adl.org/online-hate-2020

they are especially crucial in the context of social media platforms.

Examples of textual threats include toxic contents (Kim et al, 2021; Kajla et al, 2020), fake news/disinformation (de Cock Buning, 2018; Armano et al, 2018) and bullying (Grigg, 2010).

However, content is not only limited to text but can also appear in form of image or video data, as for example is dominant on platforms like Instagram and TikTok. Such user-created video and image content might convey any sort of message (verbally, non-verbally, textually or by other visual means) which can be the source of a range of threats on social media. Concrete examples are the propagation of beauty stereotypes via image data (Verrastro et al, 2020) or hyper-realistic videos/images showing people saying and doing things that never happened (Westerlund, 2019; Bursic et al, 2021), so called "deep fakes". Figure 3 Image sources [3]) demonstrates how images can be hard to distinguish between real and fake. In general, they can be misleading due to aspects like manipulation or because of the missing context of the event depicted.

Fig. 3 Real but potentially misleading images (**A** and **C**) and DeepFake/manipulated images (**B** and **D**)[3].

Given the importance of this category of threats, much research is focused on the development of dedicated detection systems as we will discuss in Section 4.

2.2 Algorithmic Threats

Besides the content itself, additional threats are caused by automatic algorithms that are used on social media platforms. These lead to the selective exposure of digital media users to news sources (Schmidt et al, 2017), risking to form closed-group polarised structures; e.g. so-called 'filter bubbles' (Nikolov et al, 2015; Geschke et al, 2019) and 'echo chambers' (Del Vicario et al, 2016; Gillani et al, 2018). Another undesired network condition is gerrymandering (Stewart et al, 2019), where users are exposed to unbalanced neighbourhood configurations. Especially in decision making framework, such as election, gerrymandering can overturn the decision of networks' participants biasing the outcome of a vote, such that one "party" wins up to 60 percent of the time in simulated elections of two-party situations where the opposing groups are equally popular through this selective presentation. This phenomenon highlight the relevance of network structure and information exposure in decision making setting.

2.3 Dynamics-induced Threats

Another type of threat is dynamics on social media, induced by the extended and fast-paced interaction between algorithms, common social tendencies and stakeholders' interests (Anderson and McLaren, 2012; Milano et al, 2021). This may lead to an escalating acceptance of toxic beliefs (Neubaum and Krämer, 2017; Stewart et al, 2019) and thus making the users' opinion susceptible to phenomena such as the diffusion of hateful content. In addition, these types of threats can lead to large-scale outbreaks of fake news (Del Vicario et al, 2016; Webb et al, 2016).

2.4 Cognitive and Socio-emotional Threats

A substantial body of work on analyzing the mechanisms of content propagation on social media exists. However, modeling the effects of the users' emotional and cognitive states as well as traits on the propagating of malicious content remains a major challenge. This is especially the case considering the significant contribution of their cognitive limits (Pennycook and Rand, 2018; Allcott and Gentzkow, 2017).

[3](A) https://www.theguardian.com/us-news/2019/apr/25/joe-biden-2020-public-gaffes-mistakes-history, (B) https://thisclimatedoesnotexist.com/, (C) https://ritzherald.com/greening-the-gray-fighting-floods-with-restoration-versus-riprap/ (D) https://www.bufale.net/bufala-la-foto-di-hillary-clinton-e-osama-bin-laden/

Such cognitive factors refer to the users' limited attention and error-prone information processing (Weng et al, 2012) that may be worsened by the emotional features of the messages (Kramer et al, 2014; Brady et al, 2017). Moreover, the lack of non-verbal communication and limited social presence (Gunawardena, 1995; Rourke et al, 1999) lead to carelessness and misbehavior as the users perceive themselves as anonymous (Diener et al, 1980; Postmes and Spears, 1998). Consequently, they do not feel judged or exposed (Whittaker and Kowalski, 2015) and deindividualize themselves and others (Lowry et al, 2016).

Another recently recognized threat in this category is *digital addiction* (Almourad et al, 2020; Nakayama and Higuchi, 2015) and it has several harmful consequences, such as unconscious and hasty user actions (Ali et al, 2015; Alrobai et al, 2016). Some of them are especially relevant for teenagers affecting their school performance and mood (Aboujaoude et al, 2006). In the last few years, it became clear that recognizing addiction to social media cannot only be based on the "connection time" criterion but also on how people behave (Taymur et al, 2016; Musetti and Corsano, 2018). As with other behavioral addictions, a crucial role may be played by the environmental structure (Ognibene et al, 2019; Kato et al, 2022).

2.5 Limited Social Media Literacy

Finally, the common lack of digital literacy among teenagers (Meyers et al, 2013) has a strong impact on the escalation of other threats, for example by favoring the spread of content-based threats and engaging in toxic dynamics (Wineburg et al, 2016). This underlines the need for education of young people in dealing with social media threats and demonstrates that automatic tools to support users in their behavior on such platforms are very important.

Teenagers also show over-reliance on algorithmic recommendations and a lack of awareness of the unwitting use of toxic content. This results in a reduction of their ability to make choices and leads towards an increasingly dangerous behavior (Banker and Khetani, 2019; Walker, 2016).

3 Related Work

The effort of supporting users on social media aims at helping them make the right decision for themselves and other people using such platforms. Strategies developed in the context of behavioral and cognitive sciences offer a well-founded framework to address these issues. In particular, nudging (Thaler and Sunstein, 2009) and boosting (Hertwig and Grüne-Yanoff, 2017) can be considered as two paradigms that have both been developed to minimize risk and harm. They do this in a way that makes use of behavioral patterns and is as unintrusive as possible, something particularly important in contexts like social media.

Nudging (Thaler and Sunstein, 2009) is a behavioral-public-policy approach aiming to push people towards more beneficial decisions through the "choice architecture" of people's environment (e.g., default settings). In a way, the machine learning-based recommender systems integrated into the social media platform already define a choice architecture that reduces the amount of content the users have to interact with, however, such recommendations are not aimed at improving users' choices in terms of collective wellbeing (Ognibene et al, 2019).

Some approaches have exploited machine learning tools to support user interactions with social media. Kyza et al (2021) propose a solution based on a web browser plugin that would use AI to support citizens dealing with misinformation by showing measures of tweets' credibility and employing a nudging mechanism that blurs out low-credibility tweets according to user's preferences. While their study uses a fact-checked dataset, it shows that such an AI-based tool may deter social media users from liking and spreading misinformation. Another work (Aprin et al, 2022) proposes a browser plugin to extend Instagram with the result of inverse image search algorithms to help users contextualize and detect fake images.

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. Zimmerman et al (2020)).

While nudges are particularly suitable for integration in social media interfaces as they may not add additional cognitive load on the users, their limitation 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 (Hertwig and Grüne-Yanoff, 2017).

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 toward 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 aware of it. 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 (Kozyreva et al, 2020; Lorenz-Spreen et al, 2020).

In particular, we refer to the concept of "media literacy" that Aufderheide (2018) defines as: the "ability of a citizen to access, analyze, and produce information for specific outcomes". Several definitions have been proposed in the literature highlighting the importance of critically approaching the media also in the light of the propagation of fake news and other toxic content as well as the influence that media can have on other citizens (Valtonen et al, 2019; Bulger and Davison, 2018).

While in this paper we present a multi-modal approach leveraging machine learning methodologies to support users and their education, algorithms and automation have taken control of many media processes such as content generation, recommendation, and filtering. Today, algorithms and machine learning are used for tracking user profiling, targeted advertising, and behaviour engineering. They have played a role in the dissemination of disinformation and misinformation as well as in impacting political opinion. The need to understand algorithm-based media requires new educational methodologies. In particular, Valtonen et al (2019) points out the necessity of combining media literacy with computing education specific to these mechanisms to allow users to cope with the changing media landscape, and Chiang and Yin (2022a) noted interactivity is a positive factor that influences the efficacy of digital media literacy.

For example, it is important to find methodologies to explain and educate about how machine learning components affect our decisions directly or by shaping our choice and information architecture, in particular in social contexts (Lomonaco et al, 2022b). It is also crucial to show the limits of such algorithms and the trade-off we should consider between our and their competencies (Chiang and Yin, 2022b).

4 Threat Detectors and Content Analyzers

The great variety of social media threats (as described in Section 2) results in challenging issues and researchers are studying how to automatically identify them. One way of bringing together the community to working on solving social media threats are workshops on these topics, e.g. (Narang et al, 2022; Kumar et al, 2020). As introduced in the beginning, another way are shared tasks. Examples include hate speech detection at SemEval 2019 (Basile et al, 2019) or Evalita 2020 (Sanguinetti et al, 2020) as well as toxic comment detection at GermEval 2021 (Risch et al, 2021b) or toxic span detection at SemEval 2021 (Pavlopoulos et al, 2021).

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. We present methods and models that have been developed as part of this project and we are using them for the detection of threats in our 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.

4.1 Text-Based Detectors

With a vast amount of social media threats taking a textual form, we proceed to present text-based detectors categorized by different threats.

4.1.1 Hate Speech and Toxic Content

An approach to profiling hate speech spreaders on Twitter was submitted to CLEF2021 and features runs for multiple languages (Akomeah et al, 2021). 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 Hoffmann and Kruschwitz (2020) 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 Turban and Kruschwitz (2022). Ensembles of transformer models and the automatic augmentation of training data were proposed. Using the common 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 Wilkens and Ognibene (2021a). The participation in the HASOC (Modha et al, 2021) 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 modeling 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 Wilkens and Ognibene (2021b). This method builds on Graph Convolutional Neural Networks (GCN) exploring different edge creation strategies and one 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 against attacks (for example, word injection) (Gröndahl et al, 2018; Hosseini et al, 2017). Some recent models can generalise the task while maintaining similar results in different platforms and languages under certain conditions (Wilkens and Ognibene, 2021b). In general this is important as small changes impact the system performance making it challenging to applying these approaches in the dynamic contexts of social media.

4.1.2 Fake News and Misinformation

To detect fake news an approach that applies automatic text summarization to compress original input documents before classifying them with a transformer model was proposed. 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 (Hartl and Kruschwitz, 2022).

Other recent methods apply ensembles of different models for fake news detection with a focus on transformer models (Tran and Kruschwitz, 2021).

In general, fake news detection datasets have frequently been proposed as part of shared tasks and we use them as for example in (Tran and Kruschwitz, 2022) or (Lomonaco et al, 2022a). While Tran and Kruschwitz (2022) apply automatic text summarization, similarly as in (Hartl and Kruschwitz, 2022), and combine this information with automatic machine translation, Lomonaco et al (2022a) introduce an approach that is based on text graphs and graph attention convolution. Although submissions were very competitive, the contributions by Tran and Kruschwitz (2022) demonstrate that this approach is highly competitive as they resulted in winning the German cross-lingual fake news detection challenge at CLEF 2022 “CheckThat!” (Tran and Kruschwitz, 2022).

4.1.3 User Beliefs and Opinions

We also use models to extract user-related properties, beliefs, and opinions as well as sentiments and emotions. Inferring and interpreting human

emotions (Poria et al, 2017) includes distinguishing between sentiment analysis, the polarity of content (e.g. Gupta et al (2018); Liu et al (2017); Guo et al (2018)), and emotion recognition (e.g. Baziotis et al (2018); Ahmad et al (2020)). In comparison, opinion extraction aims at discovering users' interests and their corresponding opinions (Wang et al, 2019). 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 (Wang et al, 2014; Chen et al, 2017).

As a lot of work in this area is going on in the NLP community, we are mainly relying on methods proposed in the literature. We use models for sentiment prediction in English (Pérez et al, 2021), German (Guhr et al, 2020), Italian (Bianchi et al, 2021) and Spanish (Pérez et al, 2021). In addition, we use models for the detection of emotions in Italian (Bianchi et al, 2021), Spanish (Plaza del Arco et al, 2020) and English (Loureiro et al, 2022).

4.2 Visual Content

One way of identifying threats in image or video data is to use textual cues related to such postings, for example associated user-comments (Mathew et al, 2019), results of transcribing the audio of a video via speech-to-text models (Hernandez Urbano Jr et al, 2021; Wu and Bhandary, 2020) or by considering text located in images (Huh et al, 2018; Giachanou et al, 2020; Armano et al, 2018).

Other methods aim at operating directly on the level of the image data: regarding the threats arising from beauty stereotypes (Verrastro et al, 2020) (e.g. to learn whether someone's feed is predominantly occupied by posts of users promoting a specific body type) we have developed a body mass index (BMI) detector that is based on a convolutional neural network and partly makes use of OpenFace (Amos et al, 2016), an open source face recognition model. It identifies a person's face within an image and predicts the BMI based on this cutout.

We also provide a gender predictor (again based on OpenFace (Amos et al, 2016)), identifying the gender of people present in an image, and an object detection algorithm that makes use of YOLOv3 (Redmon et al, 2016) to get further contextual information about the setting displayed in an image, both based on convolutional neural networks. These tools provide metadata about the image that can be used as a feature for the detection of hate speech (Das et al, 2020), violent content (Dikwatta and Fernando, 2019), and other threats.

Approaches to counteract threats like the previously mentioned "deep fakes" include the usage of deep neural networks for the detection of artifacts resulting from the production of such content (for videos see for example Montserrat et al (2020); Jung et al (2020); Hernandez-Ortega et al (2020); Sun et al (2021); Boccignone et al (2022), for images see Guarnera et al (2020); Hsu et al (2020); Chang et al (2020)). Such artifacts are for example related to image blending, the environment, behavioural anomalies, as well as audiovisual synchronization issues (Mirsky and Lee, 2021).

To improve the understanding of image feature relevance for misleadingness and correlations between user characteristics and interpretations of visual content we propose a partly crowd-sourcing-based image annotation schema. The features we consider for that are inspired by criteria used by fact-checking institutions such as the IFCN network (Graves and Anderson, 2020) and include a mixture of objective and subjective concepts. For the crowd-sourcing-based annotation, we also account for annotator characteristics using different scales such as Sosu (2013); Brotherton et al (2013); Pennycook et al (2015).

4.3 Echo Chambers and Information Gerrymandering

Another function of our tool provides support for echo chamber identification and thus helps in counteracting algorithm-based social media threats as introduced in Section 2.2. As there is no standard approach for the detection of echo chambers (Minici et al, 2022) we adopt commonly used ideas to this approach. We first apply language models for topic identification to the user's feed and the timeline posts of users connected to them in a one-hop neighborhood. In addition, we run sentiment detectors on these data. If we identify a large proportion of posts with homogeneous topics and sentiment (¿ 0.85 % of considered posts) we assume this user to be located in an echo chamber, i.e. virtually surrounded by similarly-minded

people. However, note that no information is usually available on the actual feed presented to the specific user by the platform. We suppose that if the content is shared by most of their connections it will have high chances to be presented. We thus present this aggregated information on neighborhood posts to the companion users to help evaluate the quality of their feed's sources as well as have a clearer view of the presence of social media-specific phenomena such as echo chambers and filter bubbles, which are difficult to detect for the users while affecting their experience.

5 Educational Activities and Boosting

In this section, we present the educational activities integrated complementing the companion interface's nudging functionalities with a boosting side. They aim at raising users' media literacy (Valtonen et al, 2019; Jones and Mitchell, 2016). In other words, they focus on improving students' understanding of social media dynamics and underlying computational mechanisms as well as awareness of their threats, and the strategies to use them conscientiously.

5.1 Narrative Scripts

One of our educational activities adopts the integration of image classifiers within the educational approach of the narrative scripts (Hernández-Leo et al, 2021). The narrative scripts notion combines elements from computer supported collaborative learning script mechanisms and storytelling techniques within a simulated social media platform.

The integration of machine learning tools can further assist learning scenarios covering topics related to body image stereotypes, social media algorithms and filter bubbles. Specifically, students can engage with fictional scenarios explaining the functionality of machine learning algorithms and participate in games demonstrating their effect. The objective of this work is to provide a hands-on experience of how social media algorithms work.

5.2 Education about Echo Chambers

The goal of a second activity is to increase the perception of social media influence and the possible impact of the distortions produced by echo chambers and filter bubbles. We opt for a game-oriented strategy that motivates the students and gives them the opportunity to experience the consequences of information personalisation on decision-making. The game is framed as repeated estimation task where "wisdom of crowds" (Navajas et al, 2018; Lorenz et al, 2011; Becker et al, 2017) is leveraged to simulate a bias (towards the correct or wrong direction) of the information filtering system (Lorenz et al, 2011). During this activity, participants are estimating the number of dots in an image and can revise their answers once (after also providing an aggregation of other participants' answers to them).

The intuition is that direct exposition to consequences of echo chambers and filter bubbles pushes students to being more aware of these mechanisms and their effects (i.e. when biased aggregation distorts users' unbiased opinions and its explanation). Results from a first study with around 50 students (including a baseline where the estimation task's results are not shown) confirm that explaining consequences of information personalisation on their performances during the task increase the students' awareness (Lomonaco et al, 2022b).

5.3 Awareness of Model Misclassifications

To educate teenagers about limitations of machine learning models (as used in our companion), we provide a third activity including an additional web page with examples for prediction results and statistical diagrams showing the models' average performance. Our objective is to foster the students' competence in dealing with predictions made by automatic systems, generally speaking a boosting activity (Hertwig and Grüne-Yanoff, 2017).

A part of this interface can be seen in Figure 4. We plan to use it in upcoming experiments to see whether this has positive effects on the social media literacy of teenagers.

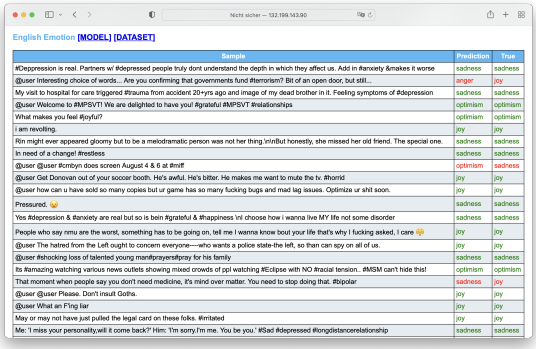

Fig. 4 Sample predictions of the English emotion prediction model for education.

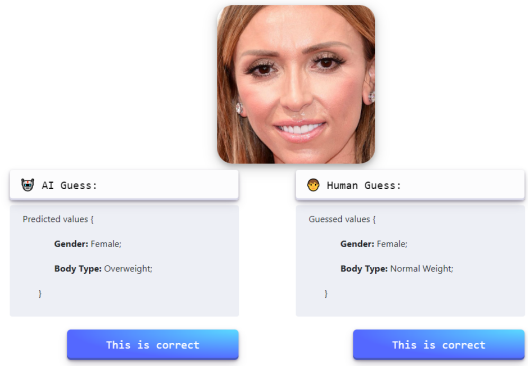

Fig. 5 Screenshot of the Trust in AI study (control condition). Participants were requested to select the prediction they trust more.

5.4 Trust in AI and Reliance on Machine Learning

A fourth activity focuses on the reliance on machine learning algorithms. We investigate the role of trust in AI and reliance on labelling systems to decorate visual content. Labelling content to signal doubtful content to reduce the spread of misinformation has been proven to be a helpful tool to increase users' capabilities to deal with fake news on social media (Gao et al, 2018; Mena, 2020) but people's trust in machine learning algorithms can also have a role in visual content misinterpretation.

We label multiple images both with the output from multiple predictors. More specifically we used the BMI, gender, and object detectors presented in Section 4.2 to label a set of images showing people. In addition, annotators were asked to annotate the same set of images with the information the models were producing.

The hypothesis is that people who trust more in AI will be more prone to rely on mislabelled content by AI. We present both sets of labels (human and AI generated) to the participants and ask them to select those that are more correct in their opinion. In the experimental condition the participants are presented the labeling methods along with the set of labels while this information is not given in the baseline condition (Figure 5).

Participants in both conditions are asked to answer a survey related to the trust in AI (Vereschak et al, 2021; McKnight et al, 2002). We plan to compare selection behavior to understand the role of trust in AI in users' image selection.

6 Conclusion

Big challenges are arising from social media usage, especially for vulnerable groups of society like teenagers, which we have summarized as part of this work. Methods for addressing these threats have been proposed and we are integrating support for multimodal content and otherwise invisible network-based threats directly into the user feed.

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. As a next step, we plan to conduct controlled user studies together with schools (in Italy, Spain and Germany) to find out how our augmented feed affects teenagers perceiving users' attitudes and content, e.g. posts, on such platforms.

In addition, several challenges remain in terms of providing efficient, extensive, and reliable machine learning-based user support tools. It is thus important to complement nudging interfaces supported by machine learning, such as our companion, with boosting educational activities to guide students in learning to leverage these tools to develop their own critical attitudes toward social media interactions instead of over relying on them.

7 Ethical Considerations

With the use of personal data and the involvement of vulnerable subjects (e.g. school children) ethical and privacy concerns arise. We strictly follow the

corresponding guidelines of our institutions (and ethical approval has been obtained before running any experiments).

We also need to stress that any individual user data (e.g. extracted from the user's social media feed) is only being used in the interaction with that specific user.

Declarations

Funding

This work was mainly supported by the project COURAGE: A Social Media Companion Safeguarding and Educating Students funded by the Volkswagen Foundation, grant number 95563, 95564, 95566, 9B145. This work has also been partially funded by the National Research Agency of the Spanish Ministry (PID2020-112584RB-C33/MICIN/AEI/10.13039/501100011033, MDM-2015-0502). D. Hernández-Leo (Serra Húnter) acknowledges the support by ICREA under the ICREA Academia programme.

Competing Interests

The authors do not have any financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

References

Aboujaoude E, Koran LM, Gamel N, et al (2006) Potential markers for problematic internet use: a telephone survey of 2,513 adults. CNS spectrums 11(10):750–755

Ahmad Z, Jindal R, Ekbal A, et al (2020) Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. Expert Systems with Applications 139:112,851

Akomeah KO, Kruschwitz U, Ludwig B (2021) University of regensburg @ pan: Profiling hate speech spreaders on twitter. In: Proceedings of the 12th Conference and Labs of the Evaluation Forum (CLEF2021). CEUR Workshop Proceedings (CEUR-WS.org), pp 2083–2089

Ali R, Jiang N, Phalp K, et al (2015) The emerging requirement for digital addiction labels. In: International working conference on requirements engineering: Foundation for software quality, Springer, pp 198–213

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236. https://doi.org/10.3386/w23089

Almourad BM, McAlaney J, Skinner T, et al (2020) Defining digital addiction: Key features from the literature. Psihologija (00):17–17

Alrobai A, McAlaney J, Phalp K, et al (2016) Online peer groups as a persuasive tool to combat digital addiction. In: International Conference on Persuasive Technology, Springer, pp 288–300

Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science

Anderson SP, McLaren J (2012) Media mergers and media bias with rational consumers. J Eur Econ Assoc 10(4):831–859

Aprin F, Chounta IA, Hoppe HU (2022) "see the image in different contexts": Using reverse image search to support the identification of fake news in instagram-like social media. In: International Conference on Intelligent Tutoring Systems, Springer, pp 264–275

Plaza del Arco FM, Strapparava C, Urena Lopez LA, et al (2020) EmoEvent: A multilingual emotion corpus based on different events. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1492–1498, URL https://aclanthology.org/2020.lrec-1.186

Armano G, Battiato S, Bennato D, et al (2018) Newsvallum: Semantics-aware text and image processing for fake news detection system. In: SEBD

Aufderheide P (2018) Media literacy: From a report of the national leadership conference on media literacy. In: Media literacy in the

information age. Routledge, p 79–86

Banker S, Khetani S (2019) 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. https://doi.org/10.1177/0743915619858057

Basile V, Bosco C, Fersini E, et al (2019) SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 54–63, https://doi.org/10.18653/v1/S19-2007, URL https://aclanthology.org/S19-2007

Baziotis C, Athanasiou N, Chronopoulou A, et al (2018) Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. arXiv preprint arXiv:180406658

Becker J, Brackbill D, Centola D (2017) Network dynamics of social influence in the wisdom of crowds. Proceedings of the national academy of sciences 114(26):E5070–E5076

Bianchi F, Nozza D, Hovy D (2021) FEEL-IT: Emotion and sentiment classification for the Italian language. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, Online, pp 76–83, URL https://aclanthology.org/2021.wassa-1.8

Bliuc AM, Faulkner N, Jakubowicz A, et al (2018) Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. Computers in Human Behavior 87:75–86

Boccignone G, Bursic S, Cuculo V, et al (2022) Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In: International Conference on Image Analysis and Processing, Springer, pp 186–195

Bowman SR, Dahl GE (2021) What will it take to fix benchmarking in natural language understanding? In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, pp 4843–4855, URL https://www.aclweb.org/anthology/2021.naacl-main.385/

Brady WJ, Wills JA, Jost JT, et al (2017) Emotion shapes the diffusion of moralized content in social networks. PNAS 114(28):7313–7318

Brotherton R, French CC, Pickering AD (2013) Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. Frontiers in psychology p 279

Bulger M, Davison P (2018) The promises, challenges, and futures of media literacy. Journal of Media Literacy Education 10(1):1–21

Bursic S, D'Amelio A, Granato M, et al (2021) A quantitative evaluation framework of video de-identification methods. In: 2020 25th international conference on pattern recognition (ICPR), IEEE, pp 6089–6095

Chan J, Ghose A, Seamans R (2016) The internet and racial hate crime: Offline spillovers from online access. MIS 40(2):381–403. https://doi.org/10.25300/MISQ/2016/40.2.05

Chang X, Wu J, Yang T, et al (2020) Deepfake face image detection based on improved vgg convolutional neural network. In: 2020 39th chinese control conference (CCC), IEEE, pp 7252–7256

Chen L, et al (2017) Building a profile of subjective well-being for social media users. PloS one 12(11)

Chiang CW, Yin M (2022a) Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In: 27th International Conference on Intelligent User Interfaces, pp 148–161

Chiang CW, Yin M (2022b) Exploring the effects of machine learning literacy interventions on

laypeople's reliance on machine learning models. Association for Computing Machinery, New York, NY, USA

Clarke B (2009) Early adolescents' use of social networking sites to maintain friendship and explore identity: implications for policy. Policy & Internet 1(1):55–89

de Cock Buning M (2018) A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation. Publications Office of the European Union

Das A, Wahi JS, Li S (2020) Detecting hate speech in multi-modal memes. arXiv preprint arXiv:201214891

Del Vicario M, Bessi A, Zollo F, et al (2016) The spreading of misinformation online. PNAS 113(3):554–559

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota

Diener E, Lusk R, DeFour D, et al (1980) Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. JPSP 39(3):449

Dikwatta U, Fernando T (2019) Violence detection in social media-review. Vidyodaya Journal of Science 22(2)

Gao M, Xiao Z, Karahalios K, et al (2018) To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. Proceedings of the ACM on Human-Computer Interaction 2(CSCW):1–16

Gerstenfeld PB, Grant DR, Chiang CP (2003) Hate online: A content analysis of extremist internet sites. ASIPP 3(1):29–44. https://doi.org/10.1111/j.1530-2415.2003.00013.x

Geschke D, Lorenz J, Holtz P (2019) 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. https://doi.org/https://doi.org/10.1111/bjso.12286

Giachanou A, Zhang G, Rosso P (2020) Multimodal multi-image fake news detection. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp 647–654

Gillani N, Yuan A, Saveski M, et al (2018) Me, my echo chamber, and i: introspection on social media polarization. In: Proceedings of the 2018 World Wide Web Conference, pp 823–831

Graves L, Anderson CW (2020) Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. New media & society 22(2):342–360

Grigg DW (2010) Cyber-aggression: Definition and concept of cyberbullying. Journal of Psychologists and Counsellors in Schools 20(2):143–156

Gröndahl T, Pajola L, Juuti M, et al (2018) All you need is: Evading hate speech detection. In: PWAIS-ACM'18, ACM, pp 2–12

Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 666–667

Guhr O, Schumann AK, Bahrmann F, et al (2020) Training a broad-coverage German sentiment classification model for dialog systems. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1627–1632, URL https://aclanthology.org/2020.lrec-1.202

Gunawardena CN (1995) Social presence theory and implications for interaction and collaborative learning in computer conferences. IJET 1(2):147–166

Guo X, Zhu B, Polanía LF, et al (2018) 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, pp 635–639

Gupta A, Agrawal D, Chauhan H, et al (2018) An attention model for group-level emotion recognition. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 611–615

Hale WC (2012) Extremism on the world wide web: A research review. Criminal Justice Studies 25(4):343–356

Hartl P, Kruschwitz U (2022) Applying automatic text summarization for fake news detection. In: Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 2702–2713

Hernandez-Ortega J, Tolosana R, Fierrez J, et al (2020) Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:201000400

Hernandez Urbano Jr R, Uy Ajero J, Legaspi Angeles A, et al (2021) A bert-based hate speech classifier from transcribed online short-form videos. In: 2021 5th International Conference on E-Society, E-Education and E-Technology, pp 186–192

Hernández-Leo D, Theophilou E, Lobo R, et al (2021) Narrative scripts embedded in social media towards empowering digital and self-protection skills. pp 394–398

Hertwig R, Grüne-Yanoff T (2017) Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science 12(6):973–986

Hoffmann J, Kruschwitz U (2020) Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings. In: EVALITA

Hosseini H, Kannan S, Zhang B, et al (2017) Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:170208138

Hsu CC, Zhuang YX, Lee CY (2020) Deep fake image detection based on pairwise learning. Applied Sciences 10(1):370

Huh M, Liu A, Owens A, et al (2018) Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV), pp 101–117

Jones LM, Mitchell KJ (2016) Defining and measuring youth digital citizenship. New media & society 18(9):2063–2079

Jung T, Kim S, Kim K (2020) Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access 8:83,144–83,154

Kajla H, Hooda J, Saini G, et al (2020) Classification of online toxic comments using machine learning algorithms. In: 2020 4th international conference on intelligent computing and control systems (ICICCS), IEEE, pp 1119–1123

Kato A, Shimomura K, Ognibene D, et al (2022) Computational models of behavioral addictions: state of the art and future directions. Addictive Behaviors p 107595

Kim JW, Guess A, Nyhan B, et al (2021) The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. Journal of Communication 71(6):922–946. https://doi.org/10.1093/joc/jqab034

Kozyreva A, Lewandowsky S, Hertwig R (2020) Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest 21(3)

Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. PNAS 111(24):8788–8790

Kumar R, Ojha AK, Lahiri B, et al (eds) (2020) Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, URL https://aclanthology.org/2020.trac-1.0

Kyza EA, Varda C, Konstantinou L, et al (2021) Social media use, trust, and technology acceptance: Investigating the effectiveness of a co-created browser plugin in mitigating the spread of misinformation on social media. In: AoIR 2021: The 22nd Annual Conference of the Association of Internet Researchers

Liu W, Wen Y, Yu Z, et al (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 212–220

Lomonaco F, Donabauer G, Siino M (2022a) Courage at checkthat! 2022: harmful tweet detection using graph neural networks and electra. Working Notes of CLEF

Lomonaco F, Ognibene D, Trianni V, et al (2022b) 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: 4th International Conference on Higher Education Learning Methodologies and Technologies Online

Lorenz J, Rauhut H, Schweitzer F, et al (2011) How social influence can undermine the wisdom of crowd effect. Proceedings of the national academy of sciences 108(22):9020–9025

Lorenz-Spreen P, Lewandowsky S, Sunstein CR, et al (2020) How behavioural sciences can promote truth, autonomy and democratic discourse online. Nature Human Behaviour

Loureiro D, Barbieri F, Neves L, et al (2022) Timelms: Diachronic language models from twitter. CoRR abs/2202.03829. URL https://arxiv.org/abs/2202.03829

Lowry PB, Zhang J, Wang C, et al (2016) 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

Mathew B, Saha P, Tharad H, et al (2019) Thou shalt not hate: Countering online hate speech. In: Proceedings of the international AAAI conference on web and social media, pp 369–380

Mathew B, Saha P, Yimam SM, et al (2021) Hatexplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence 35(17):14,867–14,875. https://doi.org/10.1609/aaai.v35i17.17745

Mcandrew FT, Jeong HS (2012) Who does what on facebook? age, sex, and relationship status as predictors of facebook use. Computers in Human Behavior 28(6):2359–2365

McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. Information systems research 13(3):334–359

Mena P (2020) Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. Policy & internet 12(2):165–183

Meyers EM, Erickson I, Small RV (2013) Digital literacy and informal learning environments: an introduction. Learning, Media and Technology 38(4):355–367. https://doi.org/10.1080/17439884.2013.783597

Milano S, Taddeo M, Floridi L (2021) Ethical aspects of multi-stakeholder recommendation systems. The Information Society 37(1):35–45

Minici M, Cinus F, Monti C, et al (2022) Cascade-based echo chamber detection. Association for Computing Machinery, New York, NY, USA

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) 54(1):1–41

Mladenović M, Ošmjanski V, Stanković SV (2021) Cyber-aggression, cyberbullying, and cybergrooming: A survey and research challenges 54(1)

Modha S, Mandl T, Shahi GK, et al (2021) 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, pp 1–3

Montserrat DM, Hao H, Yarlagadda SK, et al (2020) Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 668–669

Musetti A, Corsano P (2018) The internet is not a tool: Reappraising the model for internet-addiction disorder based on the constraints and opportunities of the digital environment. Frontiers in Psychology 9:558

Nakayama H, Higuchi S (2015) Internet addiction. Nihon rinsho Japanese journal of clinical medicine 73(9):1559–1566

Narang K, Mostafazadeh Davani A, Mathias L, et al (eds) (2022) Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), URL https://aclanthology.org/2022.woah-1.0

Navajas J, Niella T, et al (2018) Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nature Human Behaviour 2(2):126–132

Neubaum G, Krämer NC (2017) Opinion climates in social media: Blending mass and interpersonal communication. HCR 43(4):464–476

Nikolov D, Oliveira DF, Flammini A, et al (2015) Measuring online social bubbles. PeerJ Computer Science 1:e38

Ognibene D, Fiore VG, Gu X (2019) Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. Neural Networks 116:269–278. https://doi.org/https://doi.org/10.1016/j.neunet.2019.04.022

Ognibene D, Wilkens R, Taibi D, et al (2023) Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. Frontiers in Artificial Intelligence 5. https://doi.org/10.3389/frai.2022.654930, URL https://www.frontiersin.org/articles/10.3389/frai.2022.654930

Ozimek P, Baer F, Förster J (2017) Materialists on facebook: the self-regulatory role of social comparisons and the objectification of facebook friends. Heliyon 3(11):e00,449

Pavlopoulos J, Laugier L, Sorensen J, et al (2021) Semeval-2021 task 5: Toxic spans detection. Proceedings of SemEval

Pennycook G, Rand DG (2018) Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality

Pennycook G, Cheyne JA, Barr N, et al (2015) On the reception and detection of pseudo-profound bullshit. Judgment and Decision making 10(6):549–563

Pérez JM, Giudici JC, Luque FM (2021) pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. CoRR abs/2106.09462

Poria S, Cambria E, Bajpai R, et al (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37:98–125

Postmes T, Spears R (1998) Deindividuation and antinormative behavior: A meta-analysis. Psychological Bulletin 123(3):238

Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR)

Risch J, Schmidt P, Krestel R (2021a) Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Association for Computational Linguistics, Online, pp 157–163, https://doi.org/10.18653/v1/2021.woah-1.17, URL https://aclanthology.org/2021.woah-1.17

Risch J, Stoll A, Wilms L, et al (2021b) Overview of the GermEval 2021 shared task on the identification 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. Association for Computational Linguistics, Duesseldorf, Germany, pp 1–12, URL https://aclanthology.org/2021.germeval-1.1

Rourke L, Anderson T, Garrison DR, et al (1999) Assessing social presence in asynchronous text-based computer conferencing. The Journal of Distance Education/Revue de l'ducation Distance 14(2):50–71

Sanguinetti M, Comandini G, Di Nuovo E, et al (2020) Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In: EVALITA

Schafer JA (2002) Spinning the web of hate: Web-based hate propagation by extremist organizations. JCJPC

Schmidt AL, Zollo F, Del Vicario M, et al (2017) Anatomy of news consumption on facebook. PNAS 114(12):3035–3039

Shensa A, Escobar-Viera CG, Sidani JE, et al (2017) Problematic social media use and depressive symptoms among us young adults: A nationally-representative study. Social Science & Medicine 182:150–157

Shu K, Mahudeswaran D, Wang S, et al (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188. https://doi.org/10.1089/big.2020.0062, pMID: 32491943

Sosu EM (2013) The development and psychometric validation of a critical thinking disposition scale. Thinking skills and creativity 9:107–119

Stewart AJ, Mosleh M, Diakonova M, et al (2019) Information gerrymandering and undemocratic decisions. Nature 573(7772):117–121

Sun Z, Han Y, Hua Z, et al (2021) 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, pp 3609–3618

Talwar V, et al (2014) Adolescents' moral evaluations and ratings of cyberbullying: The effect of veracity and intentionality behind the event. Computers in Human Behavior 36:122–128

Tariq W, Mehboob M, Khan MA, et al (2012) The impact of social media and social networks on education and students of pakistan. International Journal of Computer Science Issues (IJCSI) 9(4):407

Taymur I, Budak E, Demirci H, et al (2016) A study of the relationship between internet addiction, psychopathology and dysfunctional beliefs. Computers in Human Behavior 61:532–536. https://doi.org/https://doi.org/10.1016/j.chb.2016.03.043

Thaler RH, Sunstein CR (2009) Nudge: Improving Decisions about Health, Wealth, and Happiness. Penguin

Tran HN, Kruschwitz U (2021) 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. Association for Computational Linguistics, Duesseldorf, Germany, pp 83–87, URL https://aclanthology.org/2021.germeval-1.12

Tran HN, Kruschwitz U (2022) 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)

Turban C, Kruschwitz U (2022) Tackling irony detection using ensemble classifiers and data augmentation. In: Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 6976–6984

Valtonen T, Tedre M, Mäkitalo K, et al (2019) Media literacy education in the age of machine learning. Journal of Media Literacy Education 11(2):20–36

Vereschak O, Bailly G, Caramiaux B (2021) How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. Proceedings of the ACM on Human-Computer Interaction 5(CSCW2):1–39

Verrastro V, Liga F, Cuzzocrea F, et al (2020) Fear the instagram: beauty stereotypes, body image and instagram use in a sample of male and female adolescents. Qwerty-Open and Interdisciplinary Journal of Technology, Culture and Education 15(1):31–49

Vidgen B, Derczynski L (2021) Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLOS ONE 15(12):1–32. https://doi.org/10.1371/journal.pone.0243300, URL https://doi.org/10.1371/journal.pone.0243300

Walker KL (2016) Surrendering information through the looking glass: Transparency, trust, and protection. Journal of Public Policy & Marketing 35(1):144–158. https://doi.org/10.1509/jppm.15.020

Wang JL, Jackson LA, Gaskin J, et al (2014) The effects of social networking site (sns) use on college students' friendship and well-being. Computers in Human Behavior 37:229–236

Wang R, Zhou D, Jiang M, et al (2019) A survey on opinion mining: From stance to product aspect. IEEE Access 7:41,101–41,124

Webb H, Burnap P, Procter R, et al (2016) Digital wildfires: propagation, verification, regulation, and responsible innovation. ACM Transactions on Information Systems (TOIS) 34(3):15

Weng L, Flammini A, Vespignani A, et al (2012) Competition among memes in a world with limited attention. Scientific reports 2:335

Westerlund M (2019) The emergence of deepfake technology: A review. Technology Innovation Management Review 9(11)

Whittaker E, Kowalski RM (2015) Cyberbullying via social media. Journal of school violence 14(1):11–29

Wilkens R, Ognibene D (2021a) bicourage: ngram and syntax gcns for hate speech detection. In: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org

Wilkens RS, Ognibene D (2021b) Mb-courage@ exist: Gcn classification for sexism identification in social networks. In: IberLEF@ SEPLN, pp 420–430

Wineburg S, McGrew S, Breakstone J, et al (2016) Evaluating information: The cornerstone of civic online reasoning. SDR 8:2018

Wu CS, Bhandary U (2020) Detection of hate speech in videos using machine learning. In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, pp 585–590

Zimmerman S, Thorpe A, Chamberlain J, et al (2020) Towards search strategies for better privacy and information. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. Association for Computing Machinery, CHIIR '20, pp 124–134

Moving Beyond Benchmarks and Competitions: Towards Addressing Social Media Challenges in an Educational Context

Dimitri Ognibene[1*], Gregor Donabauer[1,2], Emily Theophilou[3], Sathya Bursic[1], Francesco Lomonaco[1], Rodrigo Wilkens[4], Davinia Hernández-Leo[3] and Udo Kruschwitz[2]

[1]Dipartimento di Psicologia, Università Milano-Bicocca, Milan, Italy.
[2]Information Science, University of Regensburg, Regensburg, Germany.
[3]Dept. of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain.
[4]Institut Langage et Communication, Université catholique de Louvain, Louvain, Belgium.

*Corresponding author(s). E-mail(s): dimitri.ognibene@unimib.it;
Contributing authors: gregor.donabauer@ur.de; emily.theophilou@upf.edu; sathya.bursic@unimib.it; f.lomonaco5@campus.unimib.it; rodrigo.wilkens@uclouvain.be; davinia.hernandez-leo@upf.edu; udo.kruschwitz@ur.de;

Abstract

Natural language processing and other areas of artificial intelligence have seen staggering progress in recent years, yet much of this is reported with reference to somewhat limited benchmark datasets. We see the deployment of these techniques in realistic use cases as the next step in this development. In particular, much progress is still needed in educational settings, which can strongly improve users' safety on social media. We present our efforts to develop multimodal machine learning algorithms to be integrated into a social media companion aimed at supporting and educating users in dealing with fake news and other social media threats. Inside the companion environment, such algorithms can automatically assess and enable users to contextualize different aspects of their social media experience. They can estimate and display different characteristics of content in supported users' feeds, such as 'fakeness' and 'sentiment', and suggest related alternatives to enrich users' perspectives. In addition, they can evaluate the opinions, attitudes, and neighbourhoods of the users and of those appearing in their feeds. The aim of the latter process is to raise users' awareness and resilience to filter bubbles and echo chambers, which are almost unnoticeable and rarely understood phenomena that may affect users' information intake unconsciously and are unexpectedly widespread. Social media environment is rapidly changing and complex. While our algorithms show state-of-the-art performance, they rely on task-specific datasets, and their reliability may decrease over time and be limited against novel threats. The negative impact of these limits may be exasperated by users' over-reliance on algorithmic tools. Therefore, companion algorithms and educational activities are meant to increase users' awareness of social media threats while exposing the limits of such algorithms. This will also provide an educational example of the limits affecting the machine-learning components of social media platforms. We aim to devise, implement and test the impact of the companion and connected educational activities in acquiring and supporting conscientious and autonomous social media usage.

Keywords: Social media, Fake news, Hate speech, Toxic content, Education, Companion

1 Introduction

Social media have become an integral part of society in recent years. Besides all the benefits this has brought, it has also uncovered a number of serious problems including the increasing speed and the number of interactions that go beyond the users' ability to monitor and understand such content, resulting in threats such as the pervasive diffusion of fake news and biased as well as toxic content such as hate speech. A common way to address such challenges is through the adoption of natural language processing powerful state-of-the-art approaches, triggered by the paradigmatic shift that the introduction of transformer-based models (such as BERT) has led to (Devlin et al, 2019). The adoption of common benchmark collections has been another major driver in this context. Several of those datasets focus on the detection of single threats (e.g. in the domain of fake news detection (Shu et al, 2020) or for hate speech detection (Mathew et al, 2021)). Others try to unify existing text data collections, e.g. for classification of toxic content (Risch et al, 2021a; Vidgen and Derczynski, 2021).

Such benchmarks have also increasingly been utilized in a growing number of shared tasks and competitions (with leaderboards), primarily led by the machine learning (ML) community. However, a lot of work in this area remains in a purely academic classification scenario and is not being put to use in a practical context. Perhaps more importantly, it has been observed that the performance levels reported for common benchmarks do not necessarily reflect how well the algorithms will work in a realistic use case as systems are often very brittle and the performance levels do not actually transfer easily to different domains, datasets or even variations of the same dataset (Bowman and Dahl, 2021).

Instead of adopting a well-controlled setting (without any real user involvement) we aim to address an actual practical use case (which does not lend itself to being modelled around existing benchmark collections). Our starting point is the observation that social media users often have a limited understanding of the platforms and their algorithms and, more importantly, the effects of their actions on others' experiences and their role in the proliferation of toxic phenomena (Valtonen et al, 2019; Kozyreva et al, 2020). We present a framework that serves as a machine-learning-based social media education tool that aims at integrating solutions to the above-mentioned problems directly in the users' social media experience (Ognibene et al, 2023)[1]. As such the user's feed is augmented automatically with additional information on the content and underlying producing social network, as can be seen in Figure 2. Machine learning is used to trigger personalized and contextualized educational experiences that rise users' awareness about social media and its threats. At the same time, autonomous evaluation is encouraged by highlighting the principles and limits of the involved algorithmic components. The ultimate objective is to educate and empower social media users. Figure 1 gives a high-level view of the educational framework we are proposing.

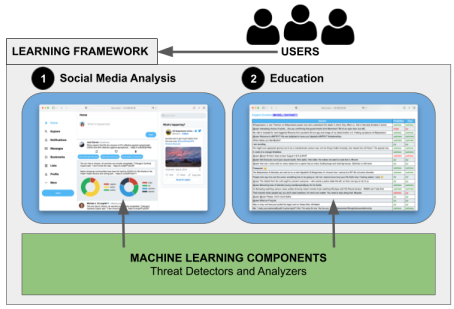

Fig. 1 Conceptual view of our proposed framework. *Social Media Analysis* shows the tool that provides additional information while browsing the feed; *Education* represents educational activities (example here: machine learning models' limitations, described in more detail in section 5.3.

In this paper, we start by discussing threats arising through social media, then present trends in how the community works on solving such issues, and then contextualize these developments in a scenario of practical use taken from the COURAGE project.

2 Social Media Threats

Threats occurring on social media cover a broad range of categories due to the vast amounts of multifaceted content on such platforms. As a result, crucial ethical and practical issues, like preserving

[1]This work is part of the COURAGE project, introducing solutions to social media harm education for teenagers (https://www.upf.edu/web/courage).

Fig. 2 Screenshot of social media content analysis results inside our Twitter demo interface. Here other user posts of the person connected to the first tweet as well as sentiment and emotion analysis are displayed. The buttons under each post allow to show/hide these additional information.

freedom of speech and allowing users to be collectively satisfied while dealing with the conflicts generated by their different opinions and contrasting interests, lead to negative influences on users and society.

Critical cases include the spread of fake news, biased content and the growing trend of hate practices (which indeed is not a new phenomenon on the internet (Gerstenfeld et al, 2003; Schafer, 2002; Chan et al, 2016)). Even though social media platforms are presenting policies against hate speech, discrimination or violent and racist content, the mentioned threats are still part of these websites[2] (Hale, 2012; Bliuc et al, 2018), underlining the need for raising awareness to the users.

Before presenting ways of how to counteract these issues in general, and how we do that with the help of our approach in the COURAGE project, we want to give a brief overview about the categories of social media threats, grouping them in **(1)** content-based, **(2)** algorithmic, **(3)** dynamics, and **(4)** cognitive and socio-emotional.

The transitions between these types of threats are fluid, making it hard to provide clear distinctions. Our focus while describing these issues lies on teenagers, which for example are heavily affected by bullying (Talwar et al, 2014; Mladenović et al, 2021), addiction (Tariq et al, 2012; Shensa et al, 2017), body stereotypes, and others (Mcandrew and Jeong, 2012; Clarke, 2009; Ozimek et al, 2017). This is also the reason why we aim at supporting this exceptionally vulnerable group of the society in the COURAGE project.

2.1 Content-Based Threats

Content-based threats are very common for all types of media, including classical outlets, but

[2]Simon Wiesenthal Center: http://www.digitalhate.net, Online Hate and Harassment Report: The American Experience 2020: https://www.adl.org/online-hate-2020

they are especially crucial in the context of social media platforms.

Examples of textual threats include toxic contents (Kim et al, 2021; Kajla et al, 2020), fake news/disinformation (de Cock Buning, 2018; Armano et al, 2018) and bullying (Grigg, 2010).

However, content is not only limited to text but can also appear in form of image or video data, as for example is dominant on platforms like Instagram and TikTok. Such user-created video and image content might convey any sort of message (verbally, non-verbally, textually or by other visual means) which can be the source of a range of threats on social media. Concrete examples are the propagation of beauty stereotypes via image data (Verrastro et al, 2020) or hyper-realistic videos/images showing people saying and doing things that never happened (Westerlund, 2019; Bursic et al, 2021), so called "deep fakes". Figure 3 Image sources [3]) demonstrates how images can be hard to distinguish between real and fake. In general, they can be misleading due to aspects like manipulation or because of the missing context of the event depicted.

Fig. 3 Real but potentially misleading images (**A** and **C**) and DeepFake/manipulated images (**B** and **D**)[3].

Given the importance of this category of threats, much research is focused on the development of dedicated detection systems as we will discuss in Section 4.

2.2 Algorithmic Threats

Besides the content itself, additional threats are caused by automatic algorithms that are used on social media platforms. These lead to the selective exposure of digital media users to news sources (Schmidt et al, 2017), risking to form closed-group polarised structures; e.g. so-called 'filter bubbles' (Nikolov et al, 2015; Geschke et al, 2019) and 'echo chambers' (Del Vicario et al, 2016; Gillani et al, 2018). Another undesired network condition is gerrymandering (Stewart et al, 2019), where users are exposed to unbalanced neighbourhood configurations. Especially in decision making framework, such as election, gerrymandering can overturn the decision of networks' participants biasing the outcome of a vote, such that one "party" wins up to 60 percent of the time in simulated elections of two-party situations where the opposing groups are equally popular through this selective presentation. This phenomenon highlight the relevance of network structure and information exposure in decision making setting.

2.3 Dynamics-induced Threats

Another type of threat is dynamics on social media, induced by the extended and fast-paced interaction between algorithms, common social tendencies and stakeholders' interests (Anderson and McLaren, 2012; Milano et al, 2021). This may lead to an escalating acceptance of toxic beliefs (Neubaum and Krämer, 2017; Stewart et al, 2019) and thus making the users' opinion susceptible to phenomena such as the diffusion of hateful content. In addition, these types of threats can lead to large-scale outbreaks of fake news (Del Vicario et al, 2016; Webb et al, 2016).

2.4 Cognitive and Socio-emotional Threats

A substantial body of work on analyzing the mechanisms of content propagation on social media exists. However, modeling the effects of the users' emotional and cognitive states as well as traits on the propagating of malicious content remains a major challenge. This is especially the case considering the significant contribution of their cognitive limits (Pennycook and Rand, 2018; Allcott and Gentzkow, 2017).

[3](A) https://www.theguardian.com/us-news/2019/apr/25/joe-biden-2020-public-gaffes-mistakes-history, (B) https://thisclimatedoesnotexist.com/, (C) https://ritzherald.com/greening-the-gray-fighting-floods-with-restoration-versus-riprap/ (D) https://www.bufale.net/bufala-la-foto-di-hillary-clinton-e-osama-bin-laden/

Such cognitive factors refer to the users' limited attention and error-prone information processing (Weng et al, 2012) that may be worsened by the emotional features of the messages (Kramer et al, 2014; Brady et al, 2017). Moreover, the lack of non-verbal communication and limited social presence (Gunawardena, 1995; Rourke et al, 1999) lead to carelessness and misbehavior as the users perceive themselves as anonymous (Diener et al, 1980; Postmes and Spears, 1998). Consequently, they do not feel judged or exposed (Whittaker and Kowalski, 2015) and deindividualize themselves and others (Lowry et al, 2016).

Another recently recognized threat in this category is *digital addiction* (Almourad et al, 2020; Nakayama and Higuchi, 2015) and it has several harmful consequences, such as unconscious and hasty user actions (Ali et al, 2015; Alrobai et al, 2016). Some of them are especially relevant for teenagers affecting their school performance and mood (Aboujaoude et al, 2006). In the last few years, it became clear that recognizing addiction to social media cannot only be based on the "connection time" criterion but also on how people behave (Taymur et al, 2016; Musetti and Corsano, 2018). As with other behavioral addictions, a crucial role may be played by the environmental structure (Ognibene et al, 2019; Kato et al, 2022).

2.5 Limited Social Media Literacy

Finally, the common lack of digital literacy among teenagers (Meyers et al, 2013) has a strong impact on the escalation of other threats, for example by favoring the spread of content-based threats and engaging in toxic dynamics (Wineburg et al, 2016). This underlines the need for education of young people in dealing with social media threats and demonstrates that automatic tools to support users in their behavior on such platforms are very important.

Teenagers also show over-reliance on algorithmic recommendations and a lack of awareness of the unwitting use of toxic content. This results in a reduction of their ability to make choices and leads towards an increasingly dangerous behavior (Banker and Khetani, 2019; Walker, 2016).

3 Related Work

The effort of supporting users on social media aims at helping them make the right decision for themselves and other people using such platforms. Strategies developed in the context of behavioral and cognitive sciences offer a well-founded framework to address these issues. In particular, nudging (Thaler and Sunstein, 2009) and boosting (Hertwig and Grüne-Yanoff, 2017) can be considered as two paradigms that have both been developed to minimize risk and harm. They do this in a way that makes use of behavioral patterns and is as unintrusive as possible, something particularly important in contexts like social media.

Nudging (Thaler and Sunstein, 2009) is a behavioral-public-policy approach aiming to push people towards more beneficial decisions through the "choice architecture" of people's environment (e.g., default settings). In a way, the machine learning-based recommender systems integrated into the social media platform already define a choice architecture that reduces the amount of content the users have to interact with, however, such recommendations are not aimed at improving users' choices in terms of collective wellbeing (Ognibene et al, 2019).

Some approaches have exploited machine learning tools to support user interactions with social media. Kyza et al (2021) propose a solution based on a web browser plugin that would use AI to support citizens dealing with misinformation by showing measures of tweets' credibility and employing a nudging mechanism that blurs out low-credibility tweets according to user's preferences. While their study uses a fact-checked dataset, it shows that such an AI-based tool may deter social media users from liking and spreading misinformation. Another work (Aprin et al, 2022) proposes a browser plugin to extend Instagram with the result of inverse image search algorithms to help users contextualize and detect fake images.

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. Zimmerman et al (2020)).

While nudges are particularly suitable for integration in social media interfaces as they may not add additional cognitive load on the users, their limitation 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 (Hertwig and Grüne-Yanoff, 2017).

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 toward 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 aware of it. 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 (Kozyreva et al, 2020; Lorenz-Spreen et al, 2020).

In particular, we refer to the concept of "media literacy" that Aufderheide (2018) defines as: the "ability of a citizen to access, analyze, and produce information for specific outcomes". Several definitions have been proposed in the literature highlighting the importance of critically approaching the media also in the light of the propagation of fake news and other toxic content as well as the influence that media can have on other citizens (Valtonen et al, 2019; Bulger and Davison, 2018).

While in this paper we present a multi-modal approach leveraging machine learning methodologies to support users and their education, algorithms and automation have taken control of many media processes such as content generation, recommendation, and filtering. Today, algorithms and machine learning are used for tracking user profiling, targeted advertising, and behaviour engineering. They have played a role in the dissemination of disinformation and misinformation as well as in impacting political opinion. The need to understand algorithm-based media requires new educational methodologies. In particular, Valtonen et al (2019) points out the necessity of combining media literacy with computing education specific to these mechanisms to allow users to cope with the changing media landscape, and Chiang and Yin (2022a) noted interactivity is a positive factor that influences the efficacy of digital media literacy.

For example, it is important to find methodologies to explain and educate about how machine learning components affect our decisions directly or by shaping our choice and information architecture, in particular in social contexts (Lomonaco et al, 2022b). It is also crucial to show the limits of such algorithms and the trade-off we should consider between our and their competencies (Chiang and Yin, 2022b).

4 Threat Detectors and Content Analyzers

The great variety of social media threats (as described in Section 2) results in challenging issues and researchers are studying how to automatically identify them. One way of bringing together the community to working on solving social media threats are workshops on these topics, e.g. (Narang et al, 2022; Kumar et al, 2020). As introduced in the beginning, another way are shared tasks. Examples include hate speech detection at SemEval 2019 (Basile et al, 2019) or Evalita 2020 (Sanguinetti et al, 2020) as well as toxic comment detection at GermEval 2021 (Risch et al, 2021b) or toxic span detection at SemEval 2021 (Pavlopoulos et al, 2021).

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. We present methods and models that have been developed as part of this project and we are using them for the detection of threats in our 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.

4.1 Text-Based Detectors

With a vast amount of social media threats taking a textual form, we proceed to present text-based detectors categorized by different threats.

4.1.1 Hate Speech and Toxic Content

An approach to profiling hate speech spreaders on Twitter was submitted to CLEF2021 and features runs for multiple languages (Akomeah et al, 2021). 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 Hoffmann and Kruschwitz (2020) 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 Turban and Kruschwitz (2022). Ensembles of transformer models and the automatic augmentation of training data were proposed. Using the common 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 Wilkens and Ognibene (2021a). The participation in the HASOC (Modha et al, 2021) 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 modeling 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 Wilkens and Ognibene (2021b). This method builds on Graph Convolutional Neural Networks (GCN) exploring different edge creation strategies and one 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 against attacks (for example, word injection) (Gröndahl et al, 2018; Hosseini et al, 2017). Some recent models can generalise the task while maintaining similar results in different platforms and languages under certain conditions (Wilkens and Ognibene, 2021b). In general this is important as small changes impact the system performance making it challenging to applying these approaches in the dynamic contexts of social media.

4.1.2 Fake News and Misinformation

To detect fake news an approach that applies automatic text summarization to compress original input documents before classifying them with a transformer model was proposed. 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 (Hartl and Kruschwitz, 2022).

Other recent methods apply ensembles of different models for fake news detection with a focus on transformer models (Tran and Kruschwitz, 2021).

In general, fake news detection datasets have frequently been proposed as part of shared tasks and we use them as for example in (Tran and Kruschwitz, 2022) or (Lomonaco et al, 2022a). While Tran and Kruschwitz (2022) apply automatic text summarization, similarly as in (Hartl and Kruschwitz, 2022), and combine this information with automatic machine translation, Lomonaco et al (2022a) introduce an approach that is based on text graphs and graph attention convolution. Although submissions were very competitive, the contributions by Tran and Kruschwitz (2022) demonstrate that this approach is highly competitive as they resulted in winning the German cross-lingual fake news detection challenge at CLEF 2022 “CheckThat!” (Tran and Kruschwitz, 2022).

4.1.3 User Beliefs and Opinions

We also use models to extract user-related properties, beliefs, and opinions as well as sentiments and emotions. Inferring and interpreting human

emotions (Poria et al, 2017) includes distinguishing between sentiment analysis, the polarity of content (e.g. Gupta et al (2018); Liu et al (2017); Guo et al (2018)), and emotion recognition (e.g. Baziotis et al (2018); Ahmad et al (2020)). In comparison, opinion extraction aims at discovering users' interests and their corresponding opinions (Wang et al, 2019). 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 (Wang et al, 2014; Chen et al, 2017).

As a lot of work in this area is going on in the NLP community, we are mainly relying on methods proposed in the literature. We use models for sentiment prediction in English (Pérez et al, 2021), German (Guhr et al, 2020), Italian (Bianchi et al, 2021) and Spanish (Pérez et al, 2021). In addition, we use models for the detection of emotions in Italian (Bianchi et al, 2021), Spanish (Plaza del Arco et al, 2020) and English (Loureiro et al, 2022).

4.2 Visual Content

One way of identifying threats in image or video data is to use textual cues related to such postings, for example associated user-comments (Mathew et al, 2019), results of transcribing the audio of a video via speech-to-text models (Hernandez Urbano Jr et al, 2021; Wu and Bhandary, 2020) or by considering text located in images (Huh et al, 2018; Giachanou et al, 2020; Armano et al, 2018).

Other methods aim at operating directly on the level of the image data: regarding the threats arising from beauty stereotypes (Verrastro et al, 2020) (e.g. to learn whether someone's feed is predominantly occupied by posts of users promoting a specific body type) we have developed a body mass index (BMI) detector that is based on a convolutional neural network and partly makes use of OpenFace (Amos et al, 2016), an open source face recognition model. It identifies a person's face within an image and predicts the BMI based on this cutout.

We also provide a gender predictor (again based on OpenFace (Amos et al, 2016)), identifying the gender of people present in an image, and an object detection algorithm that makes use of YOLOv3 (Redmon et al, 2016) to get further contextual information about the setting displayed in an image, both based on convolutional neural networks. These tools provide metadata about the image that can be used as a feature for the detection of hate speech (Das et al, 2020), violent content (Dikwatta and Fernando, 2019), and other threats.

Approaches to counteract threats like the previously mentioned "deep fakes" include the usage of deep neural networks for the detection of artifacts resulting from the production of such content (for videos see for example Montserrat et al (2020); Jung et al (2020); Hernandez-Ortega et al (2020); Sun et al (2021); Boccignone et al (2022), for images see Guarnera et al (2020); Hsu et al (2020); Chang et al (2020)). Such artifacts are for example related to image blending, the environment, behavioural anomalies, as well as audiovisual synchronization issues (Mirsky and Lee, 2021).

To improve the understanding of image feature relevance for misleadingness and correlations between user characteristics and interpretations of visual content we propose a partly crowd-sourcing-based image annotation schema. The features we consider for that are inspired by criteria used by fact-checking institutions such as the IFCN network (Graves and Anderson, 2020) and include a mixture of objective and subjective concepts. For the crowd-sourcing-based annotation, we also account for annotator characteristics using different scales such as Sosu (2013); Brotherton et al (2013); Pennycook et al (2015).

4.3 Echo Chambers and Information Gerrymandering

Another function of our tool provides support for echo chamber identification and thus helps in counteracting algorithm-based social media threats as introduced in Section 2.2. As there is no standard approach for the detection of echo chambers (Minici et al, 2022) we adopt commonly used ideas to this approach. We first apply language models for topic identification to the user's feed and the timeline posts of users connected to them in a one-hop neighborhood. In addition, we run sentiment detectors on these data. If we identify a large proportion of posts with homogeneous topics and sentiment (¿ 0.85 % of considered posts) we assume this user to be located in an echo chamber, i.e. virtually surrounded by similarly-minded

people. However, note that no information is usually available on the actual feed presented to the specific user by the platform. We suppose that if the content is shared by most of their connections it will have high chances to be presented. We thus present this aggregated information on neighborhood posts to the companion users to help evaluate the quality of their feed's sources as well as have a clearer view of the presence of social media-specific phenomena such as echo chambers and filter bubbles, which are difficult to detect for the users while affecting their experience.

5 Educational Activities and Boosting

In this section, we present the educational activities integrated complementing the companion interface's nudging functionalities with a boosting side. They aim at raising users' media literacy (Valtonen et al, 2019; Jones and Mitchell, 2016). In other words, they focus on improving students' understanding of social media dynamics and underlying computational mechanisms as well as awareness of their threats, and the strategies to use them conscientiously.

5.1 Narrative Scripts

One of our educational activities adopts the integration of image classifiers within the educational approach of the narrative scripts (Hernández-Leo et al, 2021). The narrative scripts notion combines elements from computer supported collaborative learning script mechanisms and storytelling techniques within a simulated social media platform.

The integration of machine learning tools can further assist learning scenarios covering topics related to body image stereotypes, social media algorithms and filter bubbles. Specifically, students can engage with fictional scenarios explaining the functionality of machine learning algorithms and participate in games demonstrating their effect. The objective of this work is to provide a hands-on experience of how social media algorithms work.

5.2 Education about Echo Chambers

The goal of a second activity is to increase the perception of social media influence and the possible impact of the distortions produced by echo chambers and filter bubbles. We opt for a game-oriented strategy that motivates the students and gives them the opportunity to experience the consequences of information personalisation on decision-making. The game is framed as repeated estimation task where "wisdom of crowds" (Navajas et al, 2018; Lorenz et al, 2011; Becker et al, 2017) is leveraged to simulate a bias (towards the correct or wrong direction) of the information filtering system (Lorenz et al, 2011). During this activity, participants are estimating the number of dots in an image and can revise their answers once (after also providing an aggregation of other participants' answers to them).

The intuition is that direct exposition to consequences of echo chambers and filter bubbles pushes students to being more aware of these mechanisms and their effects (i.e. when biased aggregation distorts users' unbiased opinions and its explanation). Results from a first study with around 50 students (including a baseline where the estimation task's results are not shown) confirm that explaining consequences of information personalisation on their performances during the task increase the students' awareness (Lomonaco et al, 2022b).

5.3 Awareness of Model Misclassifications

To educate teenagers about limitations of machine learning models (as used in our companion), we provide a third activity including an additional web page with examples for prediction results and statistical diagrams showing the models' average performance. Our objective is to foster the students' competence in dealing with predictions made by automatic systems, generally speaking a boosting activity (Hertwig and Grüne-Yanoff, 2017).

A part of this interface can be seen in Figure 4. We plan to use it in upcoming experiments to see whether this has positive effects on the social media literacy of teenagers.

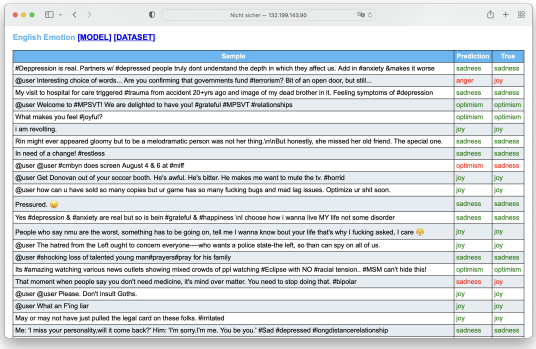

Fig. 4 Sample predictions of the English emotion prediction model for education.

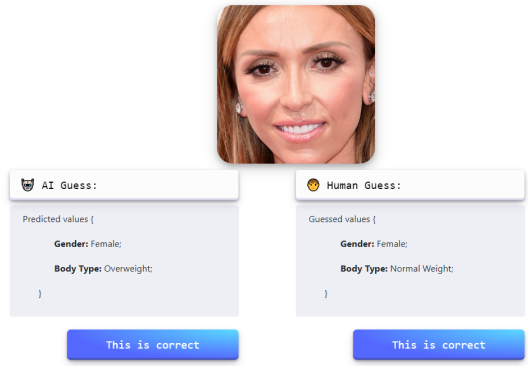

Fig. 5 Screenshot of the Trust in AI study (control condition). Participants were requested to select the prediction they trust more.

5.4 Trust in AI and Reliance on Machine Learning

A fourth activity focuses on the reliance on machine learning algorithms. We investigate the role of trust in AI and reliance on labelling systems to decorate visual content. Labelling content to signal doubtful content to reduce the spread of misinformation has been proven to be a helpful tool to increase users' capabilities to deal with fake news on social media (Gao et al, 2018; Mena, 2020) but people's trust in machine learning algorithms can also have a role in visual content misinterpretation.

We label multiple images both with the output from multiple predictors. More specifically we used the BMI, gender, and object detectors presented in Section 4.2 to label a set of images showing people. In addition, annotators were asked to annotate the same set of images with the information the models were producing.

The hypothesis is that people who trust more in AI will be more prone to rely on mislabelled content by AI. We present both sets of labels (human and AI generated) to the participants and ask them to select those that are more correct in their opinion. In the experimental condition the participants are presented the labeling methods along with the set of labels while this information is not given in the baseline condition (Figure 5).

Participants in both conditions are asked to answer a survey related to the trust in AI (Vereschak et al, 2021; McKnight et al, 2002). We plan to compare selection behavior to understand the role of trust in AI in users' image selection.

6 Conclusion

Big challenges are arising from social media usage, especially for vulnerable groups of society like teenagers, which we have summarized as part of this work. Methods for addressing these threats have been proposed and we are integrating support for multimodal content and otherwise invisible network-based threats directly into the user feed.

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. As a next step, we plan to conduct controlled user studies together with schools (in Italy, Spain and Germany) to find out how our augmented feed affects teenagers perceiving users' attitudes and content, e.g. posts, on such platforms.

In addition, several challenges remain in terms of providing efficient, extensive, and reliable machine learning-based user support tools. It is thus important to complement nudging interfaces supported by machine learning, such as our companion, with boosting educational activities to guide students in learning to leverage these tools to develop their own critical attitudes toward social media interactions instead of over relying on them.

7 Ethical Considerations

With the use of personal data and the involvement of vulnerable subjects (e.g. school children) ethical and privacy concerns arise. We strictly follow the

corresponding guidelines of our institutions (and ethical approval has been obtained before running any experiments).

We also need to stress that any individual user data (e.g. extracted from the user's social media feed) is only being used in the interaction with that specific user.

Declarations

Funding

This work was mainly supported by the project COURAGE: A Social Media Companion Safeguarding and Educating Students funded by the Volkswagen Foundation, grant number 95563, 95564, 95566, 9B145. This work has also been partially funded by the National Research Agency of the Spanish Ministry (PID2020-112584RB-C33/MICIN/AEI/10.13039/501100011033, MDM-2015-0502). D. Hernández-Leo (Serra Húnter) acknowledges the support by ICREA under the ICREA Academia programme.

Competing Interests

The authors do not have any financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

References

Aboujaoude E, Koran LM, Gamel N, et al (2006) Potential markers for problematic internet use: a telephone survey of 2,513 adults. CNS spectrums 11(10):750–755

Ahmad Z, Jindal R, Ekbal A, et al (2020) Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. Expert Systems with Applications 139:112,851

Akomeah KO, Kruschwitz U, Ludwig B (2021) University of regensburg @ pan: Profiling hate speech spreaders on twitter. In: Proceedings of the 12th Conference and Labs of the Evaluation Forum (CLEF2021). CEUR Workshop Proceedings (CEUR-WS.org), pp 2083–2089

Ali R, Jiang N, Phalp K, et al (2015) The emerging requirement for digital addiction labels. In: International working conference on requirements engineering: Foundation for software quality, Springer, pp 198–213

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236. https://doi.org/10.3386/w23089

Almourad BM, McAlaney J, Skinner T, et al (2020) Defining digital addiction: Key features from the literature. Psihologija (00):17–17

Alrobai A, McAlaney J, Phalp K, et al (2016) Online peer groups as a persuasive tool to combat digital addiction. In: International Conference on Persuasive Technology, Springer, pp 288–300

Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science

Anderson SP, McLaren J (2012) Media mergers and media bias with rational consumers. J Eur Econ Assoc 10(4):831–859

Aprin F, Chounta IA, Hoppe HU (2022) "see the image in different contexts": Using reverse image search to support the identification of fake news in instagram-like social media. In: International Conference on Intelligent Tutoring Systems, Springer, pp 264–275

Plaza del Arco FM, Strapparava C, Urena Lopez LA, et al (2020) EmoEvent: A multilingual emotion corpus based on different events. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1492–1498, URL https://aclanthology.org/2020.lrec-1.186

Armano G, Battiato S, Bennato D, et al (2018) Newsvallum: Semantics-aware text and image processing for fake news detection system. In: SEBD

Aufderheide P (2018) Media literacy: From a report of the national leadership conference on media literacy. In: Media literacy in the

information age. Routledge, p 79–86

Banker S, Khetani S (2019) 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. https://doi.org/10.1177/0743915619858057

Basile V, Bosco C, Fersini E, et al (2019) SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 54–63, https://doi.org/10.18653/v1/S19-2007, URL https://aclanthology.org/S19-2007

Baziotis C, Athanasiou N, Chronopoulou A, et al (2018) Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. arXiv preprint arXiv:180406658

Becker J, Brackbill D, Centola D (2017) Network dynamics of social influence in the wisdom of crowds. Proceedings of the national academy of sciences 114(26):E5070–E5076

Bianchi F, Nozza D, Hovy D (2021) FEEL-IT: Emotion and sentiment classification for the Italian language. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, Online, pp 76–83, URL https://aclanthology.org/2021.wassa-1.8

Bliuc AM, Faulkner N, Jakubowicz A, et al (2018) Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. Computers in Human Behavior 87:75–86

Boccignone G, Bursic S, Cuculo V, et al (2022) Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In: International Conference on Image Analysis and Processing, Springer, pp 186–195

Bowman SR, Dahl GE (2021) What will it take to fix benchmarking in natural language understanding? In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, pp 4843–4855, URL https://www.aclweb.org/anthology/2021.naacl-main.385/

Brady WJ, Wills JA, Jost JT, et al (2017) Emotion shapes the diffusion of moralized content in social networks. PNAS 114(28):7313–7318

Brotherton R, French CC, Pickering AD (2013) Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. Frontiers in psychology p 279

Bulger M, Davison P (2018) The promises, challenges, and futures of media literacy. Journal of Media Literacy Education 10(1):1–21

Bursic S, D'Amelio A, Granato M, et al (2021) A quantitative evaluation framework of video de-identification methods. In: 2020 25th international conference on pattern recognition (ICPR), IEEE, pp 6089–6095

Chan J, Ghose A, Seamans R (2016) The internet and racial hate crime: Offline spillovers from online access. MIS 40(2):381–403. https://doi.org/10.25300/MISQ/2016/40.2.05

Chang X, Wu J, Yang T, et al (2020) Deepfake face image detection based on improved vgg convolutional neural network. In: 2020 39th chinese control conference (CCC), IEEE, pp 7252–7256

Chen L, et al (2017) Building a profile of subjective well-being for social media users. PloS one 12(11)

Chiang CW, Yin M (2022a) Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In: 27th International Conference on Intelligent User Interfaces, pp 148–161

Chiang CW, Yin M (2022b) Exploring the effects of machine learning literacy interventions on

laypeople's reliance on machine learning models. Association for Computing Machinery, New York, NY, USA

Clarke B (2009) Early adolescents' use of social networking sites to maintain friendship and explore identity: implications for policy. Policy & Internet 1(1):55–89

de Cock Buning M (2018) A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation. Publications Office of the European Union

Das A, Wahi JS, Li S (2020) Detecting hate speech in multi-modal memes. arXiv preprint arXiv:201214891

Del Vicario M, Bessi A, Zollo F, et al (2016) The spreading of misinformation online. PNAS 113(3):554–559

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota

Diener E, Lusk R, DeFour D, et al (1980) Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. JPSP 39(3):449

Dikwatta U, Fernando T (2019) Violence detection in social media-review. Vidyodaya Journal of Science 22(2)

Gao M, Xiao Z, Karahalios K, et al (2018) To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. Proceedings of the ACM on Human-Computer Interaction 2(CSCW):1–16

Gerstenfeld PB, Grant DR, Chiang CP (2003) Hate online: A content analysis of extremist internet sites. ASIPP 3(1):29–44. https://doi.org/10.1111/j.1530-2415.2003.00013.x

Geschke D, Lorenz J, Holtz P (2019) 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. https://doi.org/https://doi.org/10.1111/bjso.12286

Giachanou A, Zhang G, Rosso P (2020) Multimodal multi-image fake news detection. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp 647–654

Gillani N, Yuan A, Saveski M, et al (2018) Me, my echo chamber, and i: introspection on social media polarization. In: Proceedings of the 2018 World Wide Web Conference, pp 823–831

Graves L, Anderson CW (2020) Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. New media & society 22(2):342–360

Grigg DW (2010) Cyber-aggression: Definition and concept of cyberbullying. Journal of Psychologists and Counsellors in Schools 20(2):143–156

Gröndahl T, Pajola L, Juuti M, et al (2018) All you need is: Evading hate speech detection. In: PWAIS-ACM'18, ACM, pp 2–12

Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 666–667

Guhr O, Schumann AK, Bahrmann F, et al (2020) Training a broad-coverage German sentiment classification model for dialog systems. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1627–1632, URL https://aclanthology.org/2020.lrec-1.202

Gunawardena CN (1995) Social presence theory and implications for interaction and collaborative learning in computer conferences. IJET 1(2):147–166

Guo X, Zhu B, Polanía LF, et al (2018) 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, pp 635–639

Gupta A, Agrawal D, Chauhan H, et al (2018) An attention model for group-level emotion recognition. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 611–615

Hale WC (2012) Extremism on the world wide web: A research review. Criminal Justice Studies 25(4):343–356

Hartl P, Kruschwitz U (2022) Applying automatic text summarization for fake news detection. In: Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 2702–2713

Hernandez-Ortega J, Tolosana R, Fierrez J, et al (2020) Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:201000400

Hernandez Urbano Jr R, Uy Ajero J, Legaspi Angeles A, et al (2021) A bert-based hate speech classifier from transcribed online short-form videos. In: 2021 5th International Conference on E-Society, E-Education and E-Technology, pp 186–192

Hernández-Leo D, Theophilou E, Lobo R, et al (2021) Narrative scripts embedded in social media towards empowering digital and self-protection skills. pp 394–398

Hertwig R, Grüne-Yanoff T (2017) Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science 12(6):973–986

Hoffmann J, Kruschwitz U (2020) Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings. In: EVALITA

Hosseini H, Kannan S, Zhang B, et al (2017) Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:170208138

Hsu CC, Zhuang YX, Lee CY (2020) Deep fake image detection based on pairwise learning. Applied Sciences 10(1):370

Huh M, Liu A, Owens A, et al (2018) Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV), pp 101–117

Jones LM, Mitchell KJ (2016) Defining and measuring youth digital citizenship. New media & society 18(9):2063–2079

Jung T, Kim S, Kim K (2020) Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access 8:83,144–83,154

Kajla H, Hooda J, Saini G, et al (2020) Classification of online toxic comments using machine learning algorithms. In: 2020 4th international conference on intelligent computing and control systems (ICICCS), IEEE, pp 1119–1123

Kato A, Shimomura K, Ognibene D, et al (2022) Computational models of behavioral addictions: state of the art and future directions. Addictive Behaviors p 107595

Kim JW, Guess A, Nyhan B, et al (2021) The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. Journal of Communication 71(6):922–946. https://doi.org/10.1093/joc/jqab034

Kozyreva A, Lewandowsky S, Hertwig R (2020) Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest 21(3)

Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. PNAS 111(24):8788–8790

Kumar R, Ojha AK, Lahiri B, et al (eds) (2020) Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, URL https://aclanthology.org/2020.trac-1.0

Kyza EA, Varda C, Konstantinou L, et al (2021) Social media use, trust, and technology acceptance: Investigating the effectiveness of a co-created browser plugin in mitigating the spread of misinformation on social media. In: AoIR 2021: The 22nd Annual Conference of the Association of Internet Researchers

Liu W, Wen Y, Yu Z, et al (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 212–220

Lomonaco F, Donabauer G, Siino M (2022a) Courage at checkthat! 2022: harmful tweet detection using graph neural networks and electra. Working Notes of CLEF

Lomonaco F, Ognibene D, Trianni V, et al (2022b) 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: 4th International Conference on Higher Education Learning Methodologies and Technologies Online

Lorenz J, Rauhut H, Schweitzer F, et al (2011) How social influence can undermine the wisdom of crowd effect. Proceedings of the national academy of sciences 108(22):9020–9025

Lorenz-Spreen P, Lewandowsky S, Sunstein CR, et al (2020) How behavioural sciences can promote truth, autonomy and democratic discourse online. Nature Human Behaviour

Loureiro D, Barbieri F, Neves L, et al (2022) Timelms: Diachronic language models from twitter. CoRR abs/2202.03829. URL https://arxiv.org/abs/2202.03829

Lowry PB, Zhang J, Wang C, et al (2016) 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

Mathew B, Saha P, Tharad H, et al (2019) Thou shalt not hate: Countering online hate speech. In: Proceedings of the international AAAI conference on web and social media, pp 369–380

Mathew B, Saha P, Yimam SM, et al (2021) Hatexplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence 35(17):14,867–14,875. https://doi.org/10.1609/aaai.v35i17.17745

Mcandrew FT, Jeong HS (2012) Who does what on facebook? age, sex, and relationship status as predictors of facebook use. Computers in Human Behavior 28(6):2359–2365

McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. Information systems research 13(3):334–359

Mena P (2020) Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. Policy & internet 12(2):165–183

Meyers EM, Erickson I, Small RV (2013) Digital literacy and informal learning environments: an introduction. Learning, Media and Technology 38(4):355–367. https://doi.org/10.1080/17439884.2013.783597

Milano S, Taddeo M, Floridi L (2021) Ethical aspects of multi-stakeholder recommendation systems. The Information Society 37(1):35–45

Minici M, Cinus F, Monti C, et al (2022) Cascade-based echo chamber detection. Association for Computing Machinery, New York, NY, USA

Mirsky Y, Lee W (2021) The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) 54(1):1–41

Mladenović M, Ošmjanski V, Stanković SV (2021) Cyber-aggression, cyberbullying, and cybergrooming: A survey and research challenges 54(1)

Modha S, Mandl T, Shahi GK, et al (2021) 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, pp 1–3

Montserrat DM, Hao H, Yarlagadda SK, et al (2020) Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 668–669

Musetti A, Corsano P (2018) The internet is not a tool: Reappraising the model for internet-addiction disorder based on the constraints and opportunities of the digital environment. Frontiers in Psychology 9:558

Nakayama H, Higuchi S (2015) Internet addiction. Nihon rinsho Japanese journal of clinical medicine 73(9):1559–1566

Narang K, Mostafazadeh Davani A, Mathias L, et al (eds) (2022) Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), URL https://aclanthology.org/2022.woah-1.0

Navajas J, Niella T, et al (2018) Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nature Human Behaviour 2(2):126–132

Neubaum G, Krämer NC (2017) Opinion climates in social media: Blending mass and interpersonal communication. HCR 43(4):464–476

Nikolov D, Oliveira DF, Flammini A, et al (2015) Measuring online social bubbles. PeerJ Computer Science 1:e38

Ognibene D, Fiore VG, Gu X (2019) Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. Neural Networks 116:269–278. https://doi.org/https://doi.org/10.1016/j.neunet.2019.04.022

Ognibene D, Wilkens R, Taibi D, et al (2023) Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. Frontiers in Artificial Intelligence 5. https://doi.org/10.3389/frai.2022.654930, URL https://www.frontiersin.org/articles/10.3389/frai.2022.654930

Ozimek P, Baer F, Förster J (2017) Materialists on facebook: the self-regulatory role of social comparisons and the objectification of facebook friends. Heliyon 3(11):e00,449

Pavlopoulos J, Laugier L, Sorensen J, et al (2021) Semeval-2021 task 5: Toxic spans detection. Proceedings of SemEval

Pennycook G, Rand DG (2018) Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality

Pennycook G, Cheyne JA, Barr N, et al (2015) On the reception and detection of pseudo-profound bullshit. Judgment and Decision making 10(6):549–563

Pérez JM, Giudici JC, Luque FM (2021) pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. CoRR abs/2106.09462

Poria S, Cambria E, Bajpai R, et al (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37:98–125

Postmes T, Spears R (1998) Deindividuation and antinormative behavior: A meta-analysis. Psychological Bulletin 123(3):238

Redmon J, Divvala S, Girshick R, et al (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR)

Risch J, Schmidt P, Krestel R (2021a) Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Association for Computational Linguistics, Online, pp 157–163, https://doi.org/10.18653/v1/2021.woah-1.17, URL https://aclanthology.org/2021.woah-1.17

Risch J, Stoll A, Wilms L, et al (2021b) Overview of the GermEval 2021 shared task on the identification 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. Association for Computational Linguistics, Duesseldorf, Germany, pp 1–12, URL https://aclanthology.org/2021.germeval-1.1

Rourke L, Anderson T, Garrison DR, et al (1999) Assessing social presence in asynchronous text-based computer conferencing. The Journal of Distance Education/Revue de l'ducation Distance 14(2):50–71

Sanguinetti M, Comandini G, Di Nuovo E, et al (2020) Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In: EVALITA

Schafer JA (2002) Spinning the web of hate: Web-based hate propagation by extremist organizations. JCJPC

Schmidt AL, Zollo F, Del Vicario M, et al (2017) Anatomy of news consumption on facebook. PNAS 114(12):3035–3039

Shensa A, Escobar-Viera CG, Sidani JE, et al (2017) Problematic social media use and depressive symptoms among us young adults: A nationally-representative study. Social Science & Medicine 182:150–157

Shu K, Mahudeswaran D, Wang S, et al (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188. https://doi.org/10.1089/big.2020.0062, pMID: 32491943

Sosu EM (2013) The development and psychometric validation of a critical thinking disposition scale. Thinking skills and creativity 9:107–119

Stewart AJ, Mosleh M, Diakonova M, et al (2019) Information gerrymandering and undemocratic decisions. Nature 573(7772):117–121

Sun Z, Han Y, Hua Z, et al (2021) 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, pp 3609–3618

Talwar V, et al (2014) Adolescents' moral evaluations and ratings of cyberbullying: The effect of veracity and intentionality behind the event. Computers in Human Behavior 36:122–128

Tariq W, Mehboob M, Khan MA, et al (2012) The impact of social media and social networks on education and students of pakistan. International Journal of Computer Science Issues (IJCSI) 9(4):407

Taymur I, Budak E, Demirci H, et al (2016) A study of the relationship between internet addiction, psychopathology and dysfunctional beliefs. Computers in Human Behavior 61:532–536. https://doi.org/https://doi.org/10.1016/j.chb.2016.03.043

Thaler RH, Sunstein CR (2009) Nudge: Improving Decisions about Health, Wealth, and Happiness. Penguin

Tran HN, Kruschwitz U (2021) 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. Association for Computational Linguistics, Duesseldorf, Germany, pp 83–87, URL https://aclanthology.org/2021.germeval-1.12

Tran HN, Kruschwitz U (2022) 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)

Turban C, Kruschwitz U (2022) Tackling irony detection using ensemble classifiers and data augmentation. In: Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 6976–6984

Valtonen T, Tedre M, Mäkitalo K, et al (2019) Media literacy education in the age of machine learning. Journal of Media Literacy Education 11(2):20–36

Vereschak O, Bailly G, Caramiaux B (2021) How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. Proceedings of the ACM on Human-Computer Interaction 5(CSCW2):1–39

Verrastro V, Liga F, Cuzzocrea F, et al (2020) Fear the instagram: beauty stereotypes, body image and instagram use in a sample of male and female adolescents. Qwerty-Open and Interdisciplinary Journal of Technology, Culture and Education 15(1):31–49

Vidgen B, Derczynski L (2021) Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLOS ONE 15(12):1–32. https://doi.org/10.1371/journal.pone.0243300, URL https://doi.org/10.1371/journal.pone.0243300

Walker KL (2016) Surrendering information through the looking glass: Transparency, trust, and protection. Journal of Public Policy & Marketing 35(1):144–158. https://doi.org/10.1509/jppm.15.020

Wang JL, Jackson LA, Gaskin J, et al (2014) The effects of social networking site (sns) use on college students' friendship and well-being. Computers in Human Behavior 37:229–236

Wang R, Zhou D, Jiang M, et al (2019) A survey on opinion mining: From stance to product aspect. IEEE Access 7:41,101–41,124

Webb H, Burnap P, Procter R, et al (2016) Digital wildfires: propagation, verification, regulation, and responsible innovation. ACM Transactions on Information Systems (TOIS) 34(3):15

Weng L, Flammini A, Vespignani A, et al (2012) Competition among memes in a world with limited attention. Scientific reports 2:335

Westerlund M (2019) The emergence of deepfake technology: A review. Technology Innovation Management Review 9(11)

Whittaker E, Kowalski RM (2015) Cyberbullying via social media. Journal of school violence 14(1):11–29

Wilkens R, Ognibene D (2021a) bicourage: ngram and syntax gcns for hate speech detection. In: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org

Wilkens RS, Ognibene D (2021b) Mb-courage@ exist: Gcn classification for sexism identification in social networks. In: IberLEF@ SEPLN, pp 420–430

Wineburg S, McGrew S, Breakstone J, et al (2016) Evaluating information: The cornerstone of civic online reasoning. SDR 8:2018

Wu CS, Bhandary U (2020) Detection of hate speech in videos using machine learning. In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, pp 585–590

Zimmerman S, Thorpe A, Chamberlain J, et al (2020) Towards search strategies for better privacy and information. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. Association for Computing Machinery, CHIIR '20, pp 124–134