SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Informatics, Systems and Communication

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

Information Evolution Modeling, Tracking and Analyzing in Social Media Streams

Cognome / Surname Shabunina Nome / Name Ekaterina

Matricola / Registration number 798746

Tutore / Tutor: Prof. Stefania Bandini

Supervisor: Prof. Gabriella Pasi

Coordinatore / Coordinator: Prof. Stefania Bandini

ANNO ACCADEMICO / ACADEMIC YEAR 2016/2017

Università degli Studi di Milano-Bicocca,
Dipartimento di Informatica Sistemistica e Comunicazione

Information Evolution Modeling, Tracking and Analyzing in Social Media Streams

Ph.D dissertation of

Ekaterina Shabunina

Supervisor: Prof. Gabriella Pasi

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science,
2017

Abstract

Nowadays, User Generated Content is the main source of real time news and opinions on the world happenings. Social Media, which serves as an environment for the creation and spreading of User Generated Content, is, therefore, representative of our culture and constitutes a potential treasury of knowledge. Thus, analyzing the content spread in Social Media can bring unprecedented opportunities to many areas of research and application. This matter was explored from two different angles in the first phase of this PhD thesis. In particular, the first exploratory study has focused on the correlation between the performance of companies as expressed in Social Media posts and the Stock Market values movements of the same companies. Meanwhile, the second exploratory work of the first phase of this thesis analyzed different dimensions and their combinations of the information spread under a Social Media hashtag in time. The insights and conclusions highlighted by these first two studies on Social Media have served as the basis for the development of the main ideas in this PhD. As the core focus of this thesis we propose an approach for modeling, tracking and analyzing the information evolution over time in Social Media. In particular, we propose to model a Social Media stream as a text graph. A graph degeneracy technique is used to identify the temporal sequence of the core units of information streams represented by graphs. Furthermore, as the major novelty of this work, we propose a set of measures to track and evaluate quantitatively and qualitatively the evolution of information in time. An experimental evaluation on the crawled datasets from one of the most popular Social Media platforms proves the validity and applicability of the proposed approach.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In the 21st century, User Generated Content (UGC) is the main source of real time news and opinions on the world happenings. Social Media represents the environment for the UGC creation and propagation. Thus, the streams of Social Media posts present a potential source of valuable knowledge, the analysis of which can bring unprecedented opportunities to many fields of research and applications. The numerous works in the literature focused on Social Media analysis prove the importance and value of the hidden knowledge in the UGC and its effects on the world events. Thus, during the first phase of the current PhD work the main aim was to perform exploratory analyses of textual information in Social Media; this has encompassed a study of the different dimensions of the information spread and its correlation to real world happenings. The two works developed during this first phase have brought interesting and valuable insights on the unexpected paths that information may follow while being propagated between users of a Social Network, and on the effect that it has on the decisions taken in various domains. These two exploratory studies have contributed to the development of the background knowledge and notions, as well as to the motivations of the main research work performed as the core of this PhD thesis, whose goal was the information evolution modeling, tracking and analyzing in Social Media streams. The core part of the present thesis is focused

on two main issues, which simultaneously present the two main challenges of this work. Firstly, to perform the analysis of the evolution of textual information streams over time it is necessary to be able to model these streams for the further purpose of the identification and extraction of the core units of information. These core units of information have been coined as *memes* [1], and in the environment of the fast pace information spread on the World Wide Web they have been referred to as "electronic memes", a.k.a. *ememes* [2], which is how we will refer to them in this thesis work. The second part of the core work of this thesis concerns tracking and analyzing of the evolution of information over time. For this latter task, we have proposed a set of measures for the quantitative and qualitative evaluation of the identified ememes and their evolution over time.

1.2 Context and Motivations

A *meme*, as defined by Richard Dawkins [1], is a "unit of information" or a concept that spreads from person to person within a culture. An example of a meme can be a musical melody or a whole song, a catchy phrase or a word, trending news, a behavioral pattern, etc. Memes can be manifested in several ways and forms, ranging from images [3], videos [4], texts [2, 5], etc. In the latter case, the concepts/ideas behind a meme are expressed by means of sentences in natural language, and they are replicated and spread by (slight) variations of the employed vocabulary.

Nowadays, the most common way to propagate opinions and ideas on the Web is constituted by Social Media, where users can freely express themselves: in the 21st century the User Generated Content represents the voice of people. This is due to the fact that, in the Web 2.0, users do not have external control on the content they generate, while in the previous instance of the Web only professionals were allowed to publish. In particular, micro-blogging platforms boost the creation of UGC and, therefore, they constitute an important source of news and opinions on the world happenings, which diffuse at a real-time speed. Thus, they offer to both Social Media users and researchers an invaluable repository of potential information.

This social context offers a unique opportunity to analyze the content generated by millions of users from different countries and cultural environments with the aim of excerpting cultural trends and ideas. Thus, the advantages and research opportunities brought up by Web 2.0 serve as the motivation for the present PhD thesis. In particular, such enormous social study as information evolution modeling, tracking and analyzing, which is the focus of this thesis work, is feasible in the 21st century by the virtue of the present days' Social Media platforms. The fast pace UGC creation process lead to the creation of a new term *ememe* to stand for the "Internet textual meme", as defined in [2]. The identification of such *ememes* in Social Media streams with further tracking and analyzing of their evolution is the core of the present PhD thesis.

A wide variety of definitions, representations and perceptions of (e)memes in the literature (see Sections 2.2.1 - 2.2.2 for details) emphasizes the lack of consensus in this field and the absence of a unique interpretation of a (e)meme concept in the context of Social Media, thus the need to define one; this serves as the motivation of the present PhD work. In this thesis, we define an innovative approach for the identification of ememes in Social Media streams, propose two conceptual definitions of an ememe, and propose two sets of measures (for each ememe definition) for the evaluation of the quality of the extracted ememes and the quantification of the evolution of these ememes with time.

1.3 Research Objectives and Challenges

The objectives described in this section represent simultaneously the main challenges of this thesis work. The aim of the research undertaken during the presented PhD is fourfold: 1) to analyze the potential of Social Media by performing a number of exploratory studies in the context of text analysis for the evaluation of the environment for the core work of this thesis; 2) to formally represent a stream of Social Media posts; 3) to extract from a stream of Social Media posts the potential units of information, a.k.a. *ememes*; and 4) to track the evolution in time of the identified ememes in the Social Media streams.

The first phase of this thesis is concerned with the exploratory analyses of Social Media, by the virtue of two distinct works. The objective of the first work in this first phase was to analyze the correlation between UGC and the world happenings; as a case study we have considered the Stock Market. The first step of this work regarded the identification of polarities of the performance of companies as expressed in Social Media, thus the first challenge was related to the choice of the appropriate approach (i.e. Machine Learning technique) for the classification of UGC in the very particular domain of financial markets. The second step and challenge of this first work was related to the identification of the potential correlation between the classified UGC related to a set of companies and the stock values of these companies. The second work of this first phase of the current PhD thesis focused on analyzing different dimensions of information spread in Social Media. The main challenge of this second exploratory work was the identification of important dimensions of information, and the creation of a toolkit for the identification of hidden knowledge in UGC by evaluating these different dimensions and their combinations.

The core part of this thesis is focused on three main issues, which simultaneously represent the three main challenges of this work. To perform the analysis of the evolution over time of textual information streams, it is necessary to be able to model such streams. Therefore, the textual information representation step is necessary for the performance of the further complex task of the identification and extraction of the core units of information (ememes) from these information streams. Finally, the third challenge of the core work of this thesis regards the tracking and the analysis of the evolution of information over time. For this latter task, we have proposed a set of measures for the quantitative and qualitative evaluation of the identified ememes and their evolution in time.

Besides the three main aforementioned challenges of the core work of this thesis, another important issue in the field of content-centered information propagation exists, i.e. the absence of state of the art methods for a comparative evaluation of approaches aimed at tracking the evolution of information in Social Media. Therefore, for the purpose of the evaluation of the proposed approach we have crawled a number of large and diverse datasets from a popular Social Media Platform on which we have performed a comprehensive study of the performance of

the proposed methods for textual information modeling, ememes identification, their evaluation and evolution tracking.

The tasks of the core work of the present thesis share some similarities with other adjacent research fields that in last years have been introduced and studied in the literature. These include: Topic Detection and Tracking [6, 7, 8, 9], and Event Detection [10, 11]. However, it is important to underline that even though events and (e)memes can be similarly formally defined as a set of interdependent terms, the main difference relies on the fact that an event is constrained by time while the notion of a (e)meme is not. On the contrary, one of the main characteristics of a (e)meme is its persistence in time. Nonetheless, an event may give rise to a number of (e)memes related to the event itself or to some aspects of it, however this is not mandatory. Similarly, compared to the definition of a (e)meme, the notion of a topic differs due to its higher generality. A topic may encompass several (e)memes at different points in time.

One great advantage of Social Media is the vast amount of information created in real time, which simultaneously leads to Big Data challenges. Since handling Big Data issues was not the objective of the studies performed in this thesis, for the evaluation of the proposed methods we have performed the data filtering step by means of *hashtags*, which are broadly used in Social Media platforms to identify posts on a specific topic (in Section 2.3 we provide a detailed presentation of *hashtags* along with many other interoperable Social Media features and functionalities).

Moreover, since all the studies in this thesis have been performed in Social Media platforms, one important challenge emerged due to another intrinsic characteristic of UGC: Social Media is populated by specific jargons, slangs and abbreviations (especially in some specific fields, such as finance) and a large quantity of misspellings may be intentional. Thus, the data preprocessing step is of a great value in this context, presenting a great challenge in maintaining useful information whilst filtering out all the clutter.

1.4 Research Contributions

The contributions of the present PhD thesis work are strongly related to the main objectives and challenges described in Section 1.3. As mentioned earlier, this thesis work is composed of two phases: the first phase consisted of two exploratory studied on the analysis of Social Media, while the second phase contained the core work of this thesis and consisted of three distinct parts. Each of these phases and parts has produced a set of contributions.

The main contributions of this thesis can be, therefore, summarized as follows:

- Proposed and trained a Conditional Random Fields probabilistic model for the multi-class classification of polarities of performance of companies as expressed in Social Media posts. This model has achieved an average of 93% accuracy for the 8 companies in study. This is particularly interesting since the task of polarity classification is rather challenging even for a human being, and especially in such complex topical domains of finance and IT that employ very particular jargons, slangs, symbols and abbreviations. Moreover, as part of this first exploratory study we have proposed and developed a set of visual correlation analysis, regression analysis and Granger Causality tests for the study of the relation between the polarities of performance of 8 concrete companies, as expressed in Social Media posts, and the Stock Market values of such companies.

- Proposed a set of demographic, sentiment and topical dimensions of information along with the set of guideline questions for leading the analysis of any topic under a Social Media hashtag. Further on, proposed and developed a tool for the analysis of information spread under a Social Media hashtag, which is able to extract hidden knowledge and the unexpected paths of the spread of information on Social Media. This tool is based on the combined usage of open source Natural Language Processing (NLP) tools and a module that we have developed to perform statistical analysis. Performed a case study of the proposed analysis tool by applying it to a concrete hashtag, which lead to a number of unexpected insights and observations about the spread of information under a topic.

- Proposed an innovative methodology for information evolution modeling, tracking and analyzing in Social Media streams. Firstly, proposed and developed an approach for the task of modeling textual information streams in Social Network with a graph-based representation, called Graph-of-Words. Secondly, proposed and developed an approach for the identification and the extraction of the core units of information in Social Media streams, a.k.a. ememes, with an extended k-core graph degeneracy technique. Proposed two conceptual definitions and representations of an ememe: the crisp set based and the fuzzy subset based. Thirdly, proposed and developed two sets of measures, crisp and fuzzy (for each of the two representations of an ememe), for the qualitative evaluation of the identified ememes in Social Media streams and for the quantitative evaluation of the evolution of these ememes with time. Lastly, developed an extensive experimental evaluation of the methodology proposed as the core work of this PhD thesis.

- Created a number of large datasets from the popular Social Media platform Twitter. For the first exploratory study we have crawled two datasets per each of the 8 studied companies selected out of the IT sector from the Nasdaq-100: one dataset with the Twitter posts crawled in the time period of over 5 weeks in the beginning of 2014 (part of which was manually labeled for the classified training), and one dataset with the stock market values for the same time period. For the second exploratory study we have crawled a large dataset for one month time period in the spring of 2015 containing Twitter posts on the topics of gender equality. For the core work of this thesis we have created three datasets on three distinct topics: Economy, Politics, Finance, each one crawled from Twitter for the time period of 1.5 months in the spring of 2016.

1.5 List of Publications

1. Ekaterina Shabunina and Gabriella Pasi. A graph-based approach to ememes identification and tracking in social media streams. *Knowledge-Based Systems*, October 2017

2. Ekaterina Shabunina and Gabriella Pasi. Information evolution modeling and tracking in

social media. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 599–606, August 2017

3. Ekaterina Shabunina and Gabriella Pasi. Information evolution modeling and tracking: State-of-art, challenges and opportunities. In *Proceedings of the 8th Italian Information Retrieval Workshop, Lugano, Switzerland, June 05-07, 2017.*, pages 102–105, June 2017

4. Ekaterina Shabunina, Stefania Marrara, and Gabriella Pasi. An approach to analyse a hashtag-based topic thread in twitter. In *Natural Language Processing and Information Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings*, pages 350–358, 2016

5. Ekaterina Shabunina. Correlation between stock prices and polarity of companies' performance in tweets : a crf-based approach. In *The International Symposium on Web Algorithms (iSWAG)*, June 2015

1.6 Thesis Structure

Besides the Introduction presented in this chapter, the rest of the thesis is organized as follows.

In **Chapter 2** we present the field of Information Propagation (IP) and its two main lines of work - the network-centered IP and the content-centered IP. In particular, firstly, we briefly introduce the network-centered Information Propagation field, which is not the focus of the current thesis work. Furthermore, we provide a critical and detailed overview of the literature on the field of content-centered Information Propagation, the history of the (e)meme concept, and the two main challenges of this line of works on IP - textual information representation, and information evolution tracking and analyzing. Finally, we present a brief historic overview and introduce some main notions of the context of the present PhD thesis - Social Media.

Chapter 3 presents two exploratory studies on Social Media performed as the first phase of this PhD work and which set the basis for the core work of this thesis. The first work explores

the effect of the User Generated Content on the real world events using the concrete example of the correlation between polarities of companies' performance as expressed in Social Media content and the variations of these companies' stocks. The second exploratory work focuses on the analysis of the diverse dimensions of the information spread in Social Media.

In **Chapter 4** we present the first part of the core work of this thesis - the approach proposed for the task of textual information modeling and the technique proposed for the ememes extraction. We present the background information on both of the task and the state of art approaches, followed by the formal definitions and the in-detail descriptions of the full processes. As the last part we present the two proposed formal definitions of an ememe - crisp and fuzzy.

In **Chapter 5** the second part of this thesis core work is presented - the proposed approach for the tracking and analyzing of the information evolution on Social Media. In this chapter we formally define the two sets of measures proposed for the two formal definitions of an ememe - crisp and fuzzy. The purpose of these measures is the qualitative evaluation of the identified ememes in a stream of Social Media posts and the quantitative evaluation of the evolution of these ememes in time.

Chapter 6 presents the evaluation of the two-fold core work of this PhD thesis, which includes the description of the datasets created for this purpose, their preprocessing steps, the parameter setting phase, and the presentation and discussion of the results of the evaluation of both parts on each of the crawled datasets partitioned with two time window sizes.

Finally, in **Chapter 7** we present the conclusions of this PhD thesis and summarize the achievements of the developed works. Ultimately, we draw some future work directions and opportunities.

Each of the core chapters of this PhD thesis (Chapters 3 - 6) include sections on additional background notions and related literature, and are concluded with a discussion section.

Chapter 2

Literature Review and Background

Information Propagation (IP) is a Computer Science research field that aims to analyze the spread of information on the Web through time. This issue has been explored by a number of works in the literature. The majority of the proposed approaches has considered IP as a network-centered problem [17, 18] The main focus of this line of research is to study how information spreads in a network of users. Simultaneously, another line of works on IP aims to study how the content of a piece of information evolves in time [2, 19, 20]. The scope of the present thesis is this latter approach, the objective of which is the quantitative and qualitative evaluation of the evolution of a stream of information.

Nonetheless, in Section 2.1 we present a brief review of the main works in the network-centered Information Propagation. In Section 2.2 we present an extensive review of the prior works on the content-centered Information Propagation. Finally, in Section 2.3 we present some background information on the context of the current thesis - Social Media.

2.1 Network-centered Information Propagation

In this section we present a short introduction to the main works in the network-centered Information Propagation.

Most of the approaches in the network-centered IP have been related to two fundamental models [21]: the *threshold model* and the *cascade model*. In the threshold model each node in the graph is assigned some threshold of friends that have to get "infected" by information for the node to catch it as well [18, 22]. While in the cascade model each node has a probability of adopting information whenever one of his friends adopts it [17, 23, 24]. These models, their generalizations and many others have been used in the field of network-centered IP to investigate such complex tasks as influence maximization, microblogging virality maximization, communities detection, study of network dynamics, prediction of information diffusion, etc.

One of the first works to propose *threshold models* was a 1978 study in the field of sociology by Mark Granovetter [25]. Many subsequent works have further investigated this class of models, although *Linear Threshold Model* (LTM) has served as the core for many successive generalizations.

The dynamic *cascade model* for diffusion processes originated from a work in interacting particle systems from probability theory by Thomas Liggett (1985) [26] and Richard Durrett (1988) [27]. The *Independent Cascade Model* (ICM) is the conceptually simplest and the most popular model of this type, which was originally studied in the context of marketing by Jacob Goldenberg et al. [28, 29].

A few examples of these two fundamental models are the following. In [18] a Linear Influence Model is developed in which the global influence of a node is modeled on the rate of diffusion through the implicit network. In this threshold model, the number of newly infected nodes (i.e. websites, users on Social Media) at time t is modeled as a function of influences of nodes that got infected before time t. In [30] an influence maximization problem is approached with a novel algorithm based on the Independent Cascades model, that integrates the advantages of influence ranking and influence estimation methods. A cascades prediction problem is approached in [17]. While a work in [31] studies the cascading mechanism of information-sharing on Social Media by analyzing the recurrence of the cascades on long time scale.

Based on both of the above mentioned fundamental models (LTM and ICM), the authors of [32] propose a Topic-aware Independent Cascade model and a Topic-aware Linear Threshold model.

Additionally, this work proposes another propagation model, which is focused on dealing with one of the main limitations of the Independent Cascade and Linear Threshold models - the number of parameters. This model assumes that social influence depends on authoritativeness of a user in a topic, interest of a user in a topic, and relevance of an item for a topic under a topic-aware perspective, thus greatly reducing the number of parameters.

2.2 Content-centered Information Propagation

In content-centered IP the common approach is to primarily identify the core units of information, a.k.a. (e)memes, in a stream of Social Media posts. In Section 2.2.1 we present the overview of the history of (e)memes from the introduction of the term until its first interpretations in Computer Science. Subsequently, the analysis of information evolution is performed on the identified (e)memes of the studied information stream. This leads to two main challenges in content-centered IP: the formal representation of textual information (Section 2.2.2), and its evolution tracking and analyzing (Section 2.2.3).

2.2.1 (E)memes

The term *"meme"* went through a large number of variations of its definition and of its characteristics throughout the last three decades. In 1976, in his book *The Selfish Gene* [1], Richard Dawkins coined the term *meme* to mean a "unit of human cultural evolution", analogous to a gene in genetics. Dawkins defined the memes as "selfish replicators", which have the mission to survive by leaping from brain to brain in a process of imitation. Dawkins also introduces three properties that a meme should comprise to be "successful": 1) *longevity* - the length of the life span of a replicator; 2) *fecundity* - the rate at which the meme copies are created; and 3) *copy-fidelity* - the fidelity of the replicated copies to the initial meme.

In 1986, Dawkins in "The Blind Watchmaker" [33] updates his own definition of the meme as a "pattern of information". In contrast to his 1976 remark that the strength of a meme, alike a

gene, lies in the stability of it's replicas, in 1986 Dawkins emphasizes the mutation capabilities of the meme. Additionally he suggests that memes are not autonomous and require an agent to be reproduced.

In the "Viruses of the Mind" [34], 1993, Dawkins modifies his meme definition once again by drawing an analogy between memes and epidemiological viruses that spread horizontally from medium to medium at a very high speed.

Another meme theorist D. Dennett (1995) [35] argues that memes are sensitive to the cultural and environmental pressure while diffusing and mutating. Although being a supporter of the Dawkins meme's gene analogy, Dennet refers to replicators as *"parasites"*. Similarly, Olesen (2009) [36] claims that "A meme is not a true form of life like a gene. Though, it is more like a virus and in that way we [humans] are hosts to parasites. A meme relies on his host to reproduce itself".

In Computer Science, "The Selfish Gene" [1] by Dawkins and a first application of the meme notion gave origin to the so-called "memetics" (by analogy with genetics) discipline, which aims to study the mechanisms of human culture based on the Darwin's evolutionary principles, and which is a branch of evolutionary computation [37]. Memetics claims that alike a gene, which carries biological information, a meme carries cultural information.

A big contribution to the memetics theory was the 1999 book by S. Blackmore "The meme machine" [38]. In her book, Blackmore outlines: "Memes, like genes, are replicators. That is, they are information that is copied with variation and selection. Because only some of the variants survive, memes (and hence human culture) evolve. Memes are copied by imitation, teaching and other methods, and they compete for space in our memories and for the chance to be copied again."

Besides memetics, which is not the focus of the research reported in this thesis, the notion of meme has played an important role in the context of Web 2.0, in particular, in relation to Social Media. Here, in the realm of the User Generated Content, a concept or an idea can easily replicate and propagate through time.

The study conducted by Leskovec, Backstrom, and Kleinberg [20] in 2009 was the first work in Computer Science, in which short frequently quoted phrases and their slight variations were referred to as memes.

The work reported by Bordogna and Pasi in [2] (2013) was one of the pioneering contributions to addressing the issues of the definition and the conceptual representation of "Internet textual memes", to which they refer to as *ememes*. In [2] an ememe is defined as "a unit of information (idea) replicated and propagated through the Web by one or more of the Internet-based services and applications". In the current thesis work, we follow the research reported in [2] and refer to core units of information of a stream of Social Media posts as *ememes*.

Further on, in the works proposed in the Computer Science literature in the environment of Social Media, many perceptions of how the notion of meme can be interpreted and instantiated have appeared, and several interpretations of "ememes" have been provided. A synthesis of the proposed definitions is reported here below. The software system Truthy presented in [5] assumes that a meme may have a number of types such as "hashtags" and "mentions" on Twitter, URLs and the preprocessed text of the tweet itself. In [39] a Twitter hashtag is considered as an online meme and as an event. In [40] the meme notion is associated with "short snippets of text, photos, audio, or videos" that a user posts in Social Media. In [41] the term meme is used to denote sets of topically coherent tweets. In [20, 42, 43] memes are assumed to be frequently quoted phrases. A micro ontology representation of an ememe is proposed in [2]. In [19], a meme is represented as a semantic network that is composed of entity concepts with strong links.

The discovered variety of definitions, representations and perceptions of how an meme can be interpreted on the Web brings an insight on the absence of a unique interpretation of the meme concept applied to Social Media, and the need to define one, which serves as the motivation of the present PhD thesis work.

In the next Section 2.2.2, after a short description of some of the most popular approaches to the representation of textual information, we present a detailed review of the works that have used the notion of "(e)memes" and describe the different approaches that have been used to

identify them in the context of Social Media.

2.2.2 Representation of Textual Information in Social Media Streams

The first issue of the content-centered Information Propagation is related to the formal representation of the texts in Social Media streams. In the literature, there has been proposed a large number of such representations.

The most common approach to the task of textual information representation is called the *bag-of-words* model, which was first mentioned in the 1954 work by Zellig Harris [44]. This simple representation reduces a text into a set of words by disregarding syntax.

In 1975 Gerard Salton et al. have proposed a *Vector Space Model* [45]. In this model, documents are represented as vectors, where each dimension corresponds to a separate term. Thus, if a term occurs in the document, its value in the vector is non-zero. Among several diverse approaches proposed to compute these values, a.k.a. term weights, the most famous one is the *tf-idf* (term frequency - inverse document frequency) weighting scheme developed by Karen Sparck Jones in 1972 [46, 47]. The dimensionality of the vector is the number of terms in the vocabulary (i.e. the number of distinct terms in the corpus), where by terms can be interpreted words, keywords, phrases, etc.

An approach based on an n-gram bag-of-words model is called a *language model* and was introduced in 1998 by Jay Ponte and Bruce Croft [48]. A statistical language model is a probability distribution over sequences of words. This model allows to estimate the relative likelihood of different phrases in a document collection.

In 2003 David Blei et al. proposed a generative probabilistic model called *Latent Dirichlet Allocation* (LDA) [49], in which each document is generated as a mixture of topics in a low-dimensional topic space, where the continuous-valued mixture proportions are distributed as a latent Dirichlet random variable. This model basically generates an allocation of the words in a document to topics. Thus, when computing a probability of a new document, LDA generates the most likely distribution over the topics that generated the document. Therefore, with

the LDA approach a document is modeled as a collection of topics, where each topic has a probability of generating some particular word.

Nowadays, a popular approach is to represent words as vectors; this approach is called *word2vec* and it was proposed in 2013 by Tomas Mikolov et al. [50]. This is a neural network model trained to reconstruct linguistic contexts of words. The *word2vec* objective function causes words that occur in similar contexts to have similar word embeddings.

An alternative approach is to represent a text as a graph, where the nodes are the words and the edges represent meaningful relations between words. This representation is richer than the bag-of-words model and most of the other text representation approaches, because it carries additional information. In fact a graph-based representation of texts allows the exploitation of the connections between words. Moreover, appropriate weights associated with the edges can represent the strength of the relationship between words, while the weights associated with the nodes can represent the importance of the words in a text. Furthermore, graphs allow to measure some useful properties of words (centrality, connectivity, etc). Currently, a variety of works in the literature propose graph-based representations of texts; a broad review of graph-based approaches for text representations is provided in [51].

In the field of the content-centered Information Propagation, a number of different approach have been employed to represent textual information for the further purpose of identification and extraction of the core units of it, a.k.a. (e)memes.

The work reported in [2] proposed to formally represent an ememe as a micro ontology generated by posts on the blogosphere by means of an OWL schema. In particular, the first step of the approach proposed in [2] for the formal representation of an ememe is to provide a core definition of that ememe as a user input with the OWL scheme. This definition is then subsequently transformed into Boolean queries with an OWA aggregation operator for the search for the possible candidate instances of the ememe on the blogosphere.

A semi-automatic approach to memes identification is proposed in [19]. Here memes are extracted from a corpus of documents; by this approach, a concept is defined as an n-gram that

occurs in a significant number of documents in the corpus. Two concepts are considered related when they frequently co-occur in the same documents (a threshold frequency is set by the user). A semantic network is then constructed, where nodes are entity concepts and edges are relation concepts. A meme is defined as a minimal semantic network, i.e. the network obtained by removing redundant relationships between concepts. A big limitation of this approach is that it is not fully automatic, and it requires user's feedback.

The meme-tracking approach proposed in [20] assumes that memes are frequently quoted phrases and their variations. To identify memes, a phrase graph is built, where the nodes represent the quoted phrases, and the directed edges represent the decreasing edit distance between the phrases. As an application of the proposed technique, both news and blogs articles are explored.

Similar to [20], in [42] and [43] a memes study is performed on short, frequently quoted phrases and their slight variations. The work in [42] used the MemeTracker dataset from [20] with some additional filtering phases for purity of memes for the primary goal of addressing the questions of meme mutation. While in [43] a meme is identified as a cluster of Facebook posts with the following process. First of all a set of candidate memes is generated by identifying posts that have at least 100 exact copies. Each distinct variant is then shingled into overlapping 4-word-grams, creating a term frequency vector from the 4-grams. Then memes are identified by sorting the meme variants by month and by frequency. If the cosine similarity of the 4-gram vector to all prior clusters is below 0.2 then a new meme cluster is created, otherwise the meme variant is added to the existing cluster. In the latter case, the term-frequency vector of the matching cluster is updated to include the additional variant.

The software system Truthy presented in [5] assumes that a meme may be represented by several objects such as "hashtags" and "mentions" on Twitter, URLs and the preprocessed text of the tweet itself. While in [39], a Twitter hashtag is considered as an online meme and as an event.

Similarly, in [40] different types of displays of memes from the Yahoo! Meme platform are considered: short snippets of text, photos, audio, or video, tokens in URLs, etc, which are

represented as bag-of-words.

In [41] the term meme is used to denote sets of topically coherent tweets. Authors propose an approach to meme identification in Twitter, in which tweets are initially grouped into "protomemes", i.e. groups that share one instance of entities, such as a hashtag, a mention, a URL, or a phrase. The identified protomemes are then clustered based on a set of similarity measures applied to content, network and user types. The obtained clusters are assumed to be the memes.

Relying on the above mentioned literature, a meme, defined originally as a "unit of human cultural evolution" [1], can only be evaluated a posteriori by domain experts based on historical data. An automatic detection of memes is a very difficult task due to the fact that the persistence of an meme, and therefore its aliveness is strongly influenced by people. Not every viral message that has received bursty attention can be defined as a meme.

In the current thesis work we focus on detecting *potential memes*. We refer to these potential memes as "ememes" since in Social Media a viral spread of a catchy idea or an opinion can quickly become an "electronic meme", which has to be further evaluated by time and by domain experts to determine whether it is a true meme or not. Therefore, in our work we do not claim to detect true memes but we rather try to discover potential memes, a.k.a. "ememes", based on a sample of Social Media data.

2.2.3 Information Evolution Tracking and Analyzing

The second challenge in the content-centered study of information diffusion concerns the methods for measuring, evaluating and analyzing the information evolution in time.

As already mentioned in Section 2.2.1, in his book *The Selfish Gene* (1976) [1], Richard Dawkins not only coined the term *meme*, but also introduced three properties that a meme should have: 1) *longevity* - the length of the life span of a replicator; 2) *fecundity* - the rate at which the meme copies are created; and 3) *copy-fidelity* - the fidelity of the replicated copies to the initial meme. These meme properties and their definitions have served as the inspiration and as the

basis for many of the following research works on the content-centered information diffusion analysis.

The work presented in [2] proposed a set of operators in the context of OWL micro ontologies aimed at measuring some useful properties of ememes such as fidelity (the degree to which an ememe is accurately reproduced, computed as the fuzzy matching between the given blog post and the original ememe description), mutation (the difference between the maximum and the minimum fidelity among the instances of the ememe), spread (reproductive activity of an ememe, calculated as the number of instances of the ememe in the searched source), and longevity (the time duration of the ememe's life span, which is the difference between the dates of the most recent and the oldest posts that contain the ememe instance).

Similarly, in [19] three meme metrics are proposed in the context of semantic networks: longevity (alike *longevity* in [2]), fecundity (alike *spread* in [2]) and copy-fidelity (alike *fidelity* in [2]).

One of the most recent studies on memes is presented in [43]. In this large-scale work carried out on the Facebook Social Network, the focus was on the imperfections of fidelity of information being propagated. The mutation rate measure in [43] was calculated as the proportion of meme replicas which introduce new edits as opposed to creating the exact copies. The popularity of the meme was defined as the sum of the popularity of all of its variants, where the popularity of a variant is the number of copies that variant posted on Facebook.

The work in [42] presents a study on the changes introduced in quoted texts as they diffuse through time; the authors examine properties of the quoted texts variants and uncover patterns in the rate of appearance of new variants, their length, the types of changes introduced, their popularity and the type of sites that are replicating them.

The temporal patterns of variations in quoted phrases are studied in [20], by extracting the temporal threads of all blogs and news media sites that mention the meme phrase, identifying the patterns and time lags of quoting between them, as well as analyzing their change in time in the whole thread.

A meme ranking problem is introduced in [40], where the objective is to maximize the overall

network activity by selecting which memes to show to the user. A set of heuristics is proposed, which estimates the probability that memes that users post in Social Media will go viral. The proposed heuristics take into account cosine similarity of a meme to the user's profile, the similarity among the users, and the intrinsic decay of the repost probability along time.

2.3 Social Media

Due to the fact that in the present thesis work the study of the evolution of information is performed in the context of Social Media, in the following we present some background and historic information about this environment along with some useful notions.

Recently, the traditional Web has moved into a new phase, the so-called *Web 2.0*. The name of this new Web phase and concept has been coined by Tim O'Reilly in 2005 [52]. As stated in [52] this concept originated in a conference brainstorming session between O'Reilly and MediaLive International, where it has been noted that after the burst of the 2001 dot-com bubble the Web has become more important than ever and that the surviving and new companies/applications all had some characteristics in common. Thus, the main characteristics and features of the Web 2.0 emerged: a) software as a services - software licensed on a subscription basis, centrally hosted, and continuously updated (including web apps, mashups, cloud computing, etc.); b) interactivity and rich user experience - dynamic content that is responsive to user input; c) user participation - users contributions via reviews, commenting, evaluations and User Generated Content creation (social networking websites, self-publishing platforms, etc.); d) folksonomy - free collective classification and discovery of information by users (e.g. "tagging" of images, videos, websites, etc.); e) mass participation - web access of users from world wide.

Consequently, Social Media originated as an interactive Web 2.0 Internet-based applications that allowed users to transfer from the passive state of readers to active content creators. The purpose of such Web applications is to create User Generated Content (UGC), modify it, spread and share the important, interesting or funny UGC pieces. Social Media facilitates the creation of networks/communities of users in which people create their own profiles and connect with

their friends, colleagues or like-minded strangers.

The first Social Media website was *Six Degrees*, which was created in 1997 and lasted until 2001. It was named after the concept of six degrees of separation, which is an idea (originally created by a writer Frigyes Karinthy in 1929 in a short story called "Chains") that every person in the world is connected to any other person through up to only six people. This first Social Network Service allowed users to connect to other users, send messages to their first, second and third degree contacts, and see their connection to any other user on the website.

Nowadays, among the top popular Social Media platforms there are Facebook[1], Twitter[2], Pinterest[3], LinkedIn[4], Instagram[5], YouTube[6], etc. All of these Social Media platforms have different applications and services, such as sharing inspiration, or photos, or videos, or jobs, etc.; however, they all have one common purpose - to connect people and allow them to share UGC.

Most of these popular Social Media platforms share an interoperable set of useful features, e.g. *hashtags*, *user mentions*, *re-sharing*, etc. In Social Media, a *hashtag* is a word or phrase, concatenated into one word, precedented with a hash symbol ("#"). The purpose of these hashtags is an indication of the topic of a Social Media post, for example, "#politics" or "#HeForShe". Meanwhile, *user mentions* or just *mentions* are used to explicitly tag a person in a specific Social Media post, either to draw her attention to some information piece or to indicate the authorship of it. A mention of a user in a post is performed by adding that user's name precedented with "@" symbol. The instant *re-sharing* of posts is another functionality that is implemented slightly differently in various Social Media platforms, whilst serving the same purpose. In Facebook this is performed with a "share now" button, similarly, in LinkedIn a "share" button is used, while in Twitter it is performed by the means of "retweets", and in Pinterest re-sharing of a post is performed via "save" button that adds the post to one of the user's "boards", a.k.a collections, making it visible to the users who follow that board.

[1] www.facebook.com
[2] www.twitter.com
[3] www.pinterest.com
[4] www.linkedin.com
[5] www.instagram.com
[6] www.youtube.com

All of these features and functionalities not only facilitate the creation and spreading of UGC by means of user-friendliness and ease of use of Social Media platforms but they also serve as a set of valuable utilities for researchers. In particular, hashtags allow to identify and extract Social Media posts on a specific topic or related to a certain discussion thread. Meanwhile, user mentions allow to analyze the content related to a specific user. Whereas, the instant re-sharing functionality permits to analyze the paths of spreading of an exact piece of information.

Chapter 3

Social Media Exploratory Analyses

The emergence of the Web 2.0 has granted user the freedom to interact with other users and to contribute contents to the World Wide Web. The most common way to propagate opinions and ideas on the Web is constituted by Social Media, which facilitates the creation and sharing of User Generated Content (UGC). Thus, Social Media provides the possibility to analyze the content generated by a vast number of users from different countries and social backgrounds to the aim of excerpting cultural trends and ideas as those spread in real time. Consequently, it has motivated novel research directions, and it has also provided new perspectives within existing ones. The analysis of the UGC can bring insights on the behavior of users in Social Networks, the patterns of their interactions, its correlations with real world events, and the structure of the information spread depending on the phenomenon driving it.

Thus, during the first phase of the current PhD work the main focus has been on the exploratory analysis of textual information in Social Media, and its effect on the real world happenings. This phase has lead to two distinctive works focusing on different aspects of information spread on Social Media, which has served as the basis for the development of the core ideas in this thesis. The first work has aimed at studying the correlation between Social Media content and real work events, in particular, we have focused on Stock Market oscillations (Section 3.1). While the focus of the second work has been on the spread of textual information in Social Media based on different dimensions: demographic, topical and sentiment related (Section 3.2).

In Section 3.3 we present a discussion on the two performed studies and the conclusions they have lead to.

3.1 Correlation between Stock Prices and polarity of companies' performance in Tweets

The analysis of the information in microblogs like Twitter is currently receiving an increasing attention from analysts and researchers. The growth in popularity and adoption of Social Media platforms like Twitter offers an unmatched and unprecedented source of data for opinion and fact mining. The current research challenge is to exploit such data and relate them to real world events, to the purpose of predicting/estimating similar events or trends.

The first exploratory work undertaken as part of the presented PhD thesis has addressed the issue of the existence of correlation between the polarity of companies' performance as expressed in tweets and the variation of these companies' stock prices. To address this issue, firstly, we have collected tweets (over a given period) containing companies' ticker tags, or "cashtag" as they are referred to in Twitter. We have then classified the extracted tweets as negative, neutral, or positive, based on the information they contain about the performance of the companies. Finally we have studied the correlation between the various measures of performance polarity and the total tweets volume against the actual stock closing prices and the traded volumes of company securities.

Intuitively, one can assume that such correlation would be high, due to the fact that investors, traders and (most of) the other players in the financial markets are human beings. As such, they are indeed affected by different information, rumors and events along with their choice of actions on the stock market, and they tend to share them (once sharing was with close-by people, nowadays it is on Social Media). There might even be intentional usage of Social Networks by investors and traders, who use Twitter as a platform to communicate, share their knowledge and even try to influence the masses to perform trading actions in their favor.

Recently some research [53, 54, 55] have addressed the problem of correlating stock prices to Social Network sentiments. The approach we have proposed is different: in fact, we do not exploit the polarity of the opinions in the tweets, but rather the polarity of the facts about the company's performance as expressed by the Twitter users. For example, a tweet of a user declaring she is happy that Google Inc. stock price is going down, would be labeled as "negative" due to the negative performance of the company's stock rather than the "positive" sentiment of the micro-blogger. We rely on the Conditional Random Fields probabilistic model over manually labeled datasets for multi-class classification of the companies' performance. We achieved a classification accuracy of 93% for the 8 companies in the experiment. This is especially interesting if we consider that the task of polarity classification of tweets is very complex (even for a human being). Moreover, companies' stock related tweets are part of a special financial domain, which employs a very specific set of jargons and slangs, with particular abbreviations and symbols. Last but not least, company-related tweets tend to be even shorter than the 140 symbols limit for a Twitter microblog, often stating only the action on the stock market that the investor has performed or that he is going to perform. An example of such an intricate and implicit semantics is the following tweet: "short $GOOG". As one can see, it contains very little information for a classifier to build an accurate classification model.

Nevertheless, the performed regression analyses have proven that the total volume of tweets has a stronger statistically significant relationship with the stock measures than the performance polarity measures. This can be seen as the proof of the common saying that there is no such a things as bad publicity.

In the following sections we introduce the relevant background concepts (Section 3.1.1), present the proposed methodology (Section 3.1.2), the experimental scenario (Section 3.1.3) and the achieved results (Section 3.1.4). Lastly, in Section 3.1.5 we present some conclusions of this exploratory study.

3.1.1 Background

In this Section the background notions useful to understand the work reported in this chapter are presented.

3.1.1.1 Conditional Random Fields

For the task of classification in this first exploratory study we have adopted Conditional Random Fields (CRF) [56], a framework for building probabilistic models to segment and label sequence data. Due to its sequential nature, the CRF classifier performs better than the common bag-of-words approaches, especially for short texts, where bag-of-words approaches usually fail, due to the sparseness of the resulting feature vector.

CRFs have a similar structure to the Conditional Markov Model (CMM), and consequently share the same benefits of the CMM over generative models such as Hidden Markov models (HMM), but instead of using a directed graph as CMM, CRFs use an undirected graph. CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. Therefore, the label bias problem does not arise for CRFs because the weights of different states can be traded off against each other, which can lead to accuracy improvements. The results presented in [56] demonstrate that "even when the models are parameterized exactly in the same way, CRFs are more robust to inaccurate modeling assumptions than MEMMs or HMMs, and resolve the label bias problem, which affects the performance of MEMMs". This is the main reason we thought that CRF could be a successful technique in the stock market field.

3.1.1.2 Stock Market

Despite the variety of analyses that are performed by the players in the financial markets in order to determine whether to buy or sell a given security (and at which prices to do so), they can be generally divided into two main categories: technical and fundamentalist. While the first one is basically an attempt of applying mathematical models to understand the behavior

of a given security and try to forecast its future movements, the second one is based on the study of intrinsic value of the company whose shares are under consideration. The value of a company is based on its capacity of generating cash in the future. Buying a company share, in this sense, is the same as buying an expected future cash flow.

The relationship between the general ascertained performance polarity in tweets related to a given company, in this sense, may be a good proxy of its future profitability. Therefore, the quantitative performance polarity analysis of these tweets may be a good indicator of the future profitability of a company, in a way that it could be correlated to its stock performance. This relationship has been put to test in this fist exploratory work.

3.1.2 Methodology

Figure 3.1 shows the architectural overview of the proposed approach for the study of the correlation between the polarities of the companies' performance on Social Media and these companies' stock variations. The whole process is divided into five main parts: Data Pre-processing, Data processing, Training, Twitter data labeling, and Regression analysis of tweets against stock market variables. In the following we provide a detailed description of each one of these parts.

3.1.2.1 Data Pre-processing

First of all, we have crawled the tweets to be analyzed. This has been obtained through the Twitter Search API by the "cashtags" of the companies considered for the analysis. A cashtag consists of the company's ticker code with "$" sign before it, for example, "$msft" for Microsoft Corp.

For the Twitter raw data filtering process numerous text processing steps have been applied: multiple lines breaks elimination; e.g., filtering out of all the non-English data, using an off-the-shelf language detection technique[1], etc. Then, the original tweet timestamps were normalized

[1] https://code.google.com/p/language-detection/

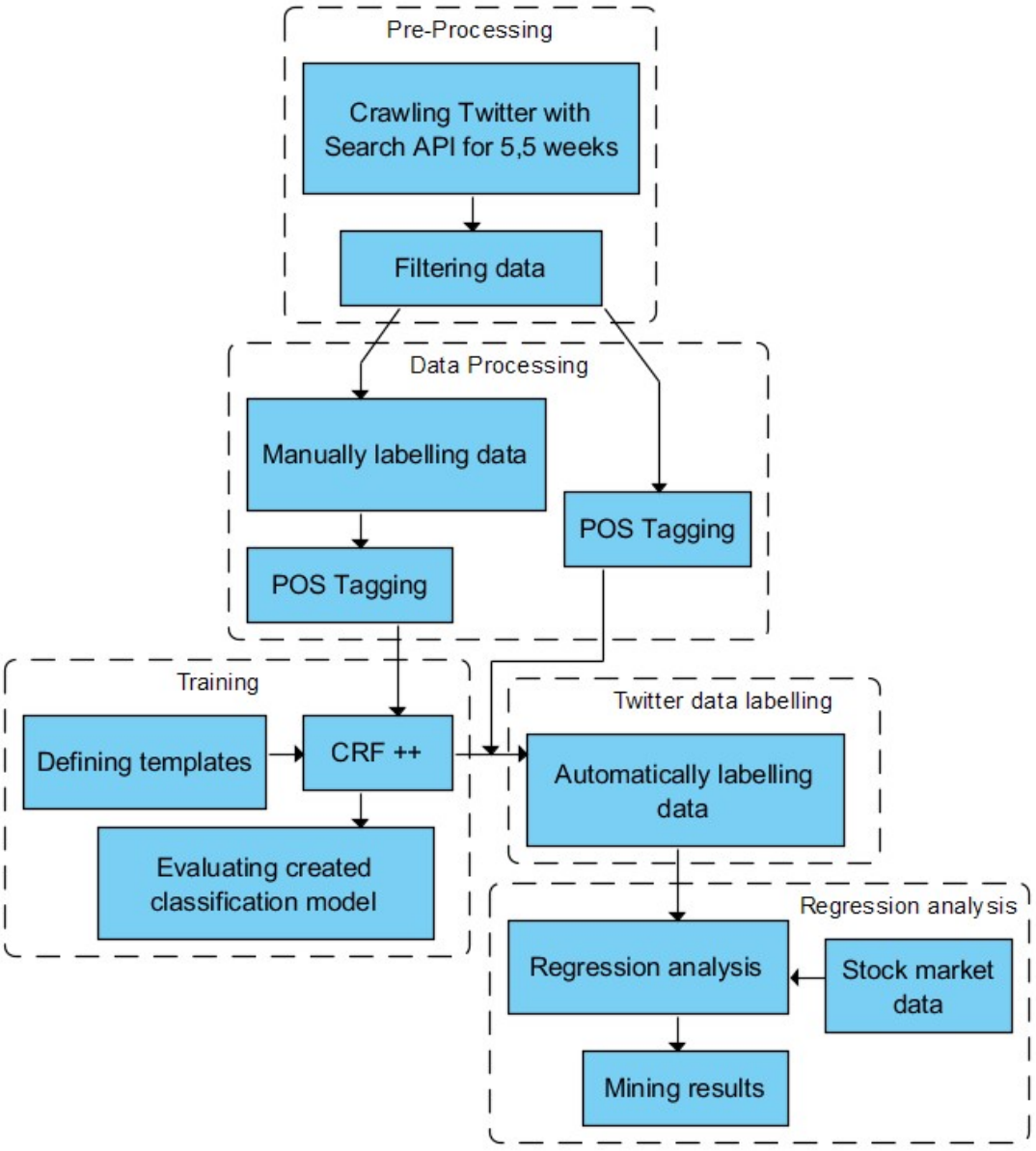

Figure 3.1: Overview of the steps applied by the proposed approach.

to the American Eastern Time zone to be comparable to the NASDAQ stock exchange time and grouped in reference to the trading hours of the Nasdaq Stock Exchange. This is based on the assumption that the tweets posted after the closing time of the stock market (16:00) have to be correlated to the next day's stock market performance, on Friday after 16:00 as well as within the weekends chatter are assumed to have a reflection on the following Monday.

3.1.2.2 Data Processing

The tweets to be used for the training of the classifier (first 20% of the collected data per each company, reported in Table 3.1) were manually labeled at tweet level in the sense of the company's performance, as either "positive", "negative" or "neutral", where "neutral" stands for the tweets with no specific indicator of the company's performance or those with a mixed one.

Company	Total	Positive	Negative	Neutral	Manually labeled
EA	2,589	366	21	2,202	250
ORCL	2,642	611	101	1,930	405
GOOG	31,529	7,184	3,216	21,128	9,177
MSFT	16,646	4,433	550	11,663	3,516
ADBE	1,399	307	102	990	265
INFY	383	100	3	280	71
LOGI	261	147	1	113	63
YHOO	9,606	2,821	886	5,898	2,191

Table 3.1: Companies Twitter dataset statistics

Each tweet has been tokenized at word level and Part-of-Speech tagged with the java "CMU ARK Twitter POS Tagger" [2].

The tweets have then been represented as an array of strings (both the training and the testing data) as required by the CRF.

3.1.2.3 Model Training, Data Labeling, and Regression Analysis

In this first exploratory study we have chosen to use a general-purpose tool CRF++ [3], a simple, customizable, and open source implementation of Conditional Random Fields for segmenting and labeling sequential data.

CRF++ has been trained with the manually labeled corpus of tweets in which each instance (tweet) is a sequence of words, labeled according to the class of the instance to which they belong.

[2] http://www.ark.cs.cmu.edu/TweetNLP/
[3] http://crfpp.googlecode.com/svn/trunk/doc/index.html

To obtain an as much unbiased estimate of the performance of the system as possible, the classification model has been trained using 10-fold cross-validation. We have performed numerous experiments with bigram templates, varying different combinations of word and POS tag features in the search for the best classifiers performance.

The Twitter data labeling phase has consisted of automatically labeling all the data that has not been manually labeled (80% of the collected data per each company), with the created classification model.

For the statistical analysis of the regression, we have employed the software Minitab 17 [4].

3.1.3 Experimental setting

For the experiments we selected 8 companies in the IT sector from the Nasdaq-100, in particular those that showcased a diverse Twitter activity distribution. They are Google Inc. Class C (GOOG), Microsoft Corp. (MSFT), Yahoo! Inc. (YHOO), Electronic Arts Inc. (EA), Adobe Systems Inc. (ADBE), Infosys Technologies (INFY), Logitech International S.A. (LOGI), and Oracle Corp. (ORCL).

For each company two datasets have been created. The first one has consisted of tweets about the company's stock market performance and has been constructed by crawling Twitter only based on the company's cashtag. The second dataset contained the company's stock closing prices and negotiated volumes for the same time period as of the Twitter dataset, and has been created using Bloomberg[5].

The experiment datasets enclose tweets over a 5 weeks' time span, from 17th of April, 2014 until 24th of May, 2014. The first week (17th of April until 24th of April, 2014) has been manually labeled and used as the training data for the classifier. Table 3.1 shows the numerical statistics related to these datasets. One important note: as clearly observable from Table 3.1, the number of negative tweets is remarkably lower than of positive tweets for all 8 companies

[4]http://www.minitab.com/enus/products/minitab/
[5]http://www.bloomberg.com/

studied.

3.1.4 Experimental Results

In this Section we present the evaluation of the classification model and the results of the statistical correlation analysis performed in our first exploratory study on Social Media.

3.1.4.1 Classifier Performance

We have conducted several experiments with different templates, by varying the involved features. The best result has been achieved with a template which included the following features: two previous words, current and two next words, their POS tags and these words combinations: word before previous word / previous word, previous/current, current/next, and next/ next after next words.

Table 3.2 presents the best achieved performance measures of the classification models for each of the 8 companies studied, selected out of the 10-folds.

C	**Total**				**Positive**				**Negative**				**Neutral**			
	Acc	Prec	Recall	F	Acc	Prec	Recall	F	Acc	Prec	Recall	F	Acc	Prec	Recall	F
EA	0.93	0.95	0.67	0.79	0.94	0.97	0.85	0.90	0.95	1	0.17	0.29	0.91	0.87	1	0.93
ORCL	0.93	0.94	0.78	0.85	0.94	0.97	0.84	0.90	0.95	1	0.50	0.67	0.91	0.86	1	0.93
GOOG	0.94	0.90	0.88	0.89	0.95	0.88	0.84	0.86	0.95	0.88	0.84	0.86	0.92	0.93	0.95	0.94
MSFT	0.92	0.88	0.81	0.84	0.90	0.88	0.85	0.86	0.96	0.88	0.63	0.74	0.90	0.89	0.94	0.91
ADBE	0.92	0.92	0.76	0.83	0.94	0.91	0.91	0.91	0.93	1	0.40	0.57	0.90	0.86	0.97	0.91
INFY	0.89	0.57	0.48	0.52	0.86	0.88	0.47	0.61	0.98	0	0	0	0.84	0.84	0.98	0.90
LOGI	0.94	0.68	0.63	0.61	0.92	1	0.89	0.94	1	0	0	0	0.92	0.75	1	0.86
YHOO	0.94	0.90	0.90	0.90	0.91	0.88	0.79	0.83	0.98	0.92	0.94	0.93	0.92	0.91	0.95	0.93

Table 3.2: Classifier's Performance for all considered companies. The first column contains the companies' ticker tags. Acc: Accuracy, Prec: Precision, Recall: Recall, F: F measure.

The average accuracy over all companies is 93% and for some companies it has even exceeded 94% (Logitech Systems Inc.: 94.4%, Google Inc.: 94.1%). While evaluating these results, it is necessary to consider, that since the experiments were conducted on real Twitter data: a) the training sets for all companies were non-symmetric; b) the training sets for some companies were rather numerically limited; c) the sets have been manually labeled with inherent subjectivity;

d) the study has been performed over the rather complex financial markets domain; e) the datasets consisted of short by nature Twitter messages and; f) as features the analysis has only used word combinations and Part-of-Speech tags.

It is relevant to note that, as mentioned in Section 3.1.3, people tend to tweet mostly positive comments about the companies' stock performance instead of negative ones, therefore the difference between the number of positive and negative tweets is very large. Due to this finding, the performance measures for the negative tweets are not very coherent to those achieved for the positive and neutral classes. For example, referring to Table 3.2, Logitech (logi) had only 1 negative tweet out of 258 for the whole 5 weeks period considered and, similarly, Infosys (infy) had 3 negative tweets out of 383. For these two companies, for the negative class the accuracy is ideal, while the recall and precision are consequently 0.

3.1.4.2 Visual correlation

Before starting to discuss the results, it is important to stress that correlation does not imply causality. This means that, for instance, observing a high number of positive tweets may not be (and is very likely not) the reason why the stock performance for the studied day was positive. On the contrary, it may simply mean that the stock variation caused a positive reaction on the Twitter users, which in their turn posted positive content about the company.

The aim of this preliminary study[6], is to visually identify whether a similar trend is present in the number of classified and total amount of tweets with the movements of the stocks' closing prices and negotiated volumes.

In this section, we firstly would like to understand if the daily variation of the stocks' closing prices presented itself as a better variable than the absolute closing price (in USD) to be related to the results obtained in terms of the difference in between the number of positively and negatively classified tweets per each trading day. Secondly, we present the comparison of the total number of tweets, disregarding the classification, as a correlation variable to two

[6]Our work is not an attempt of creating a tool for predicting stock market movements.

measures absolute closing price change (the module of the % change in a given day) and the total negotiated volume. The second analysis was to infer if the closing prices movement or the total traded volume presented better adherence measures to the total chatter on Twitter about the stocks of companies. An additional overall goal pursued by this study was the comparison of the polarity identified in the tweets and the total volume of tweets as the correlation measures to the stocks' movements.

As pointed out in Section 3.1.4.1, people tend to tweet less when the company has a negative situation. Due to this finding, it became impossible to use a logical measure for the analysis of the stock price accumulated inertia, which was infinitely increasing due to a recurrently higher number of positive tweets in the datasets.

In this section, although the charts have been plotted for all companies studied, Oracle Inc. is used as the demonstrative example.

First we looked at the two charts per each company: "Pos-Neg" (number of positive tweets minus number of negative tweets per day) and the closing prices (Figure 3.2), against "Pos-Neg" and the change in the closing prices (Figure 3.3). As becomes clearly observable, in the first figure a slight correlation takes place only sporadically, while on the second figure it is explicitly visible that the plots follow similar patterns. This finding proves the necessity of the price change measure in assessing the adherences between stocks' closing prices and the social indicators of the companies' performances in tweets.

The second visual comparison of the measures is between the total number of tweets (independently from classification) in a given day and the absolute (module) price change (Figure 3.4), against the total number of tweets and the total negotiated volume for the given security (Figure 3.5). Even if this comparison does not imply anything about the quality of the classification procedure, it is important for the understanding of the relationships between social networking activities and real world phenomena. As observable, in both figures the visual correlation is strongly present.

Figure 3.2: Visual correlation for Oracle Inc.: Pos-Neg and Closing Prices

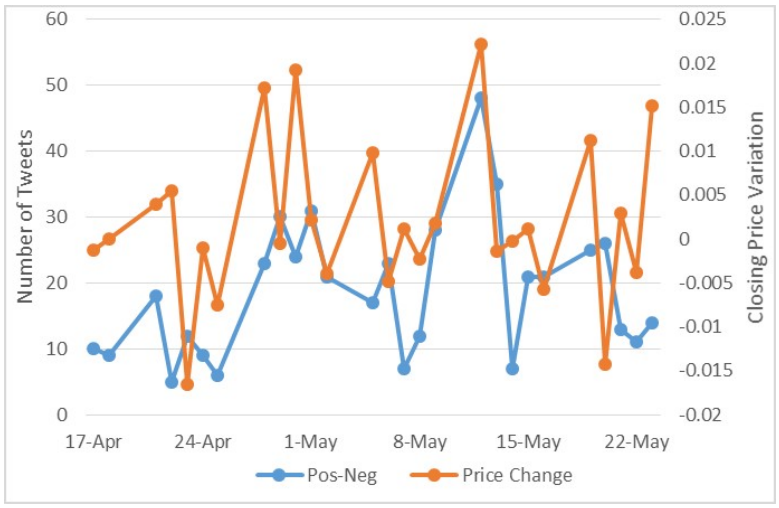

Figure 3.3: Visual correlation for Oracle Inc.: Pos-Neg and Closing Prices Change

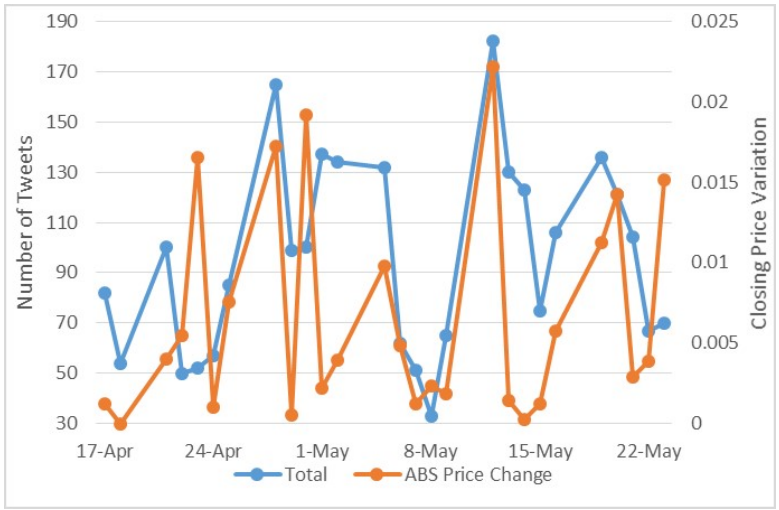

Figure 3.4: Visual correlation for Oracle Inc.: Total Tweets Volume and Absolute Closing Prices Change

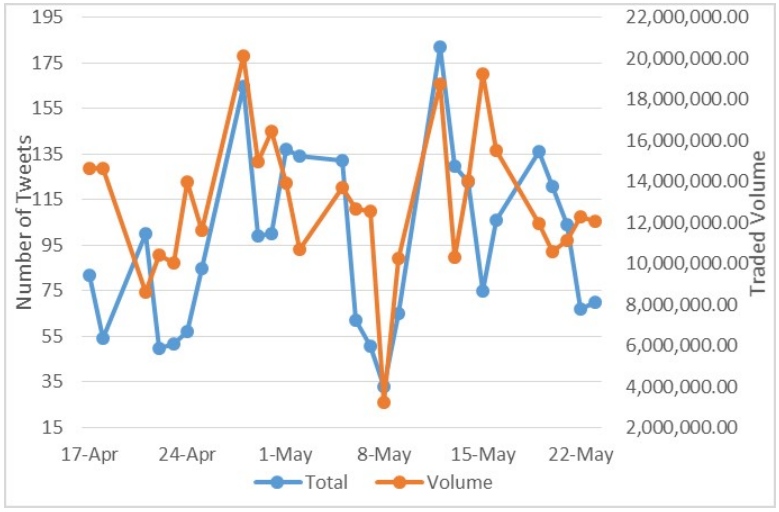

Figure 3.5: Visual correlation for Oracle Inc.: Total Tweets Volume and Total Traded Volume

3.1.4.3 Regression Analysis

The aim of this section is to go beyond the simple visual analysis of the plotted data, with the purpose of providing statistical certainty on the inferred conclusions. To formally validate the adherence of the classification results with the real world observations, a set of statistical regression analyses have been performed, using the same pairs of variables reported in the previous subsection of Section 3.1.4.2.

In order to verify the statistical significance of the relationship between two variables the ANOVA (ANalysis Of VAriance) methodology is applied to the linear and quadratic regression. The test examines the statistical significance (p-value) of the relationship between the two variables at a given threshold of 5% (p<0.05). It presents the explanatory capacity of the regression model ($R - sq$) and the correlation (r) between two variables.

For the comparison of the first two pairs, similarly to what was observable in the Figures 3.2 and 3.3, the closing price change seems to be a better correlation measure to the difference between the number of positive and negative tweets about the companies' performance. We achieved only one case of $p = 0.049$, on the edge of the given statistical significance threshold (p<0.05), for the daily performance polarity difference and closing prices. For the performance polarity difference and the price change 3 out of 8 companies had a statistically significant relationship: two of which with an ideal p<0.001, and one even with a strong regression model's explanatory

capacity of over 90% and 0.85 correlation between these two variables.

Using the total number of tweets as the independent variable and as the dependent variables both the absolute (modular) price change and the total traded volume, similar results were obtained as already noted in the previous subsection (Figures 3.4 and 3.5). 5 out of 8 relationships were statistically significant for both cases. For these statistically significant cases the average explanatory capacity was: 36% for the absolute price change case and 39% for the total traded volume case. It is worth noticing that for Electronic Arts Inc. has been observed an almost perfect $R - sq$ of 95.43% in the quadratic regression model for the absolute price change case; and an 82.28% respectively for the traded volume case.

3.1.4.4 Granger Causality

The Granger Causality test [57] is widely used in determining if one time-series $X(t)$ is useful in forecasting another time-series $Y(t)$. Since the regression analysis shows only the presence of correlation between the variables and not the causality of the relationship, the Granger tests were applied in the two possible directions for each variable at study of all the 8 companies considered for the experiment.

The obtained results for this test are rather random and did not enable us to make coherent conclusions. The cases with a causal relationship were mostly not in accordance with the regression analysis results. Though in most cases in which there is the presence of Granger causality the stocks are the cause of the tweets generation and not the other way around, as predicted. In general, only half of the companies had at least one positive Granger Causality relationship. Only for Microsoft Inc. all cases but one (total number of tweets versus the traded volume) have a causality relationship. But the stock markets variable always appears as the forecaster for the tweets measure.

3.1.5 Conclusions

The polarity classification method presented in this first exploratory study achieved an average accuracy of 93%, using as features only word combinations and Part-of-Speech tags.

The observation referring to the strongly lower amount of negative tweets for all 8 companies compared to positive or negative tweets brings to a curious conclusion: Twitter users in general tend either to post more actively tweets about the positive companies' performance or not to post much when the company has a negative event or price movement.

The explicitly visible correlation was achieved for all companies in three cases: the polarity difference and the closing prices change, the total number of tweets and the absolute by module price change, and the total number of tweets and the total negotiated volume for the given security. The regression analysis results presented 5 out of 8 statistically significant relationships for both pairs with the total tweets volume as independent variable; and 3 out of 8 for the polarity difference and price change. The polarity difference and the stock's closing price neither visually nor quantitatively present valuable relationship. Therefore, it is possible to make the strong conclusion that the polarities of companies' performance in tweets contains less information than the general volume of tweets posted, thus, making possibly unnecessary to undergo the time consuming and demanding task of tweets classification.

The Granger causality test showed that the Twitter chatter about the stocks of companies at study did not have a forecasting capacity of the stocks' prices or volumes.

Although the experimental results do not allow to extract general conclusions for the whole stock market, due to the limitation in time and number of companies analyzed, the achieved observations referring to the accuracy of the classification models and the found adherences of the polarity tweets with the stock market values show that, nevertheless, the proposed approach is promising and has a strong potential, especially if properly combined with other stock market analysis tools.

3.2 Diffusion Dimensions of Content under Twitter Hashtags

In last years, with the spread of Social Media the analysis of how content related to a given topic is spread through the Web constitutes an interesting research issue. In particular, Twitter is a microblogging platform where the discussion around a given topic can be easily identified and tracked by means of the so-called *hashtags* (e.g., #HeForShe). The assumption at the basis of this second exploratory study on Social Media is that a hashtag is representative of an *eccentric* topic, as outlined in [58]. This allows to easily crawl a thread of tweets, which can offer a good repository to perform various kinds of analysis, related to different characteristics of both the content and the users generating it. Analyses of this kind can give several interesting insights on how a certain topic diffuses based on different dimensions of information and among users, also based on their characteristics.

The objective of this second exploratory study was to propose a tool that, based on a set of standard Natural Language Processing (NLP) techniques, allows to analyze some of the most important dimensions of the content posted under a Twitter hashtag and the characteristics of the users who contributed to the hashtag spread. In particular, the aim of the proposed approach was to monitor over time the content sub-topicality and sentiment in the tracked tweets, and its relation with demographics such as gender and age. Differently to the well studied phenomenon of information diffusion, in our second exploratory work we aimed to study the *evolution of content* on a given Twitter topic in a limited timeframe.

In the following sections we present the context of the second exploratory study on Social Media (Section 3.2.1), the proposed approach in detail (Section 3.2.2), the application of our approach on a specific topic (Section 3.2.3) and, finally, the conclusions of this second exploratory study (Section 3.2.4).

3.2.1 Background

In this section we shortly review the literature that shares some common background with the approach proposed in this second exploratory study on the textual content spread in Social Media.

Several research works focus on the study of the demographics of the Social Network users who generate the information spread. The work in [59] studies the representativeness of the USA population by the Twitter users in geography, gender and race/ethnicity. In [59] the gender of the Twitter users is detected through a comparison of the first word of the self-reported name in the user's profile to the list of the most popular names reported by the USA Social Security Administration. In the research conducted in [60] the tweets from reporters of 51 US newspapers have been analyzed for the gender ratio in the quotes. The result reported in [60] is in line with most previous works on this topic, which present that women are less likely to be used as quoted sources overall. Additionally, in [42] it has been discovered that blogs are more likely to quote the source without introducing any changes, in contrast to professional journalists.

Another research line analyzes hashtags, their diffusion and the characteristics of their spread. The work in [61] proposes a hybrid approach based on linear regression, which efficiently predicts the acceptance of hashtags based on the combination of content, context, temporal and topological features of the hashtags. In [58] the authors propose a framework to capture the dynamics of hashtags based on their topicality (number of retweets), interactivity (number of replies), diversity (number of unique retweet sources), and prominence (expected number of followers). In [62] the "rich-get-richer" phenomenon is demonstrated for the hashtags distribution, showing that the frequency of use of the most popular hashtags increases faster than the popularity of the vast majority of them. This phenomenon is furthermore proven in [58]. In [63] and [18] the authors outline that different topics have distinct patterns of information spread.

The field of content-centered Information Propagation, presented in Chapter 2, has a broader

aim than the purpose of the approach proposed in this exploratory study: the content-centered IP focuses on the information modeling, analyzing and tracking, while in the study presented in this section the focus is on the different dimensions of textual information spread under a given hashtag, which practically concerns only the analysis part of the content-centered IP. Nonetheless, the basis of content-centered IP, which emphasizes the importance of the content of the information piece being spread, serves for this exploratory study as well. Thus, we claim that the analysis of the UGC over time is an important research issue. To this aim, in the second exploratory study of this first phase of the current PhD thesis we proposed a simple approach with a related tool that can easily support a few basic kinds of analysis related to both users and tweets. The proposed approach is presented in Section 3.2.2.

3.2.2 The Proposed Approach

In this section we describe the proposed approach for analyzing some dimensions of content and characteristics of the evolution of a topic (that in this work is simply identified by a hashtag in Twitter) in a given time interval.

3.2.2.1 The Twitter Topic Analysis Questions and Dimensions

In this section we present a set of questions that we proposed to use as guidelines for the analysis of any stream of tweets associated with a topic:

1. What is the main sentiment along the tweets stream? Does it change over time and w.r.t. user gender or age?
2. How are the users who contributed to the analysed topic divided w.r.t. their topics of interests?
3. Who are the main contributors to the tweets stream? Are they private users or corporates?
4. Does the topic have a goal? If so, are the statistics of the analysis in line with the main purpose of the topic creation?

5. What are the characteristics (e.g. gender, age, etc) of the leading quoted sources?

The above questions can be more specifically tailored to the topic under consideration. To answer the above questions, we have proposed a combined usage of open source NLP tools and of a module that we have developed to perform some specific statistical analyses; we refer to this combination as the Analysis Tool.

The tool, presented in the next section, works with a set of sentiment and demographic dimensions aimed at gathering the statistics necessary to perform the analyses. The considered dimensions are:

- the **gender** (male or female) of the user;
- the **age** of the user, to identify if the topic affected diverse groups of the Twitter population;
- the **user sentiment**;
- the **tweet sentiment**;
- the main **topic of interest** of each user (e.g., Arts, Business...), to identify the interests of the users who posted in the analyzed topic;
- the **quoted sources** (via the analysis of *retweets*), to detect the trusted voice of the crowd and the dominant contributors to the spread of information;
- **sub-topicality** (via terms frequency and jointly used hashtags), to analyze changes in the information carried by the given hashtag.

3.2.2.2 The Analysis Tool

In this section we describe the Analysis Tool we have developed. The tool has been created to provide all the necessary means to answer the questions reported in the previous subsection.

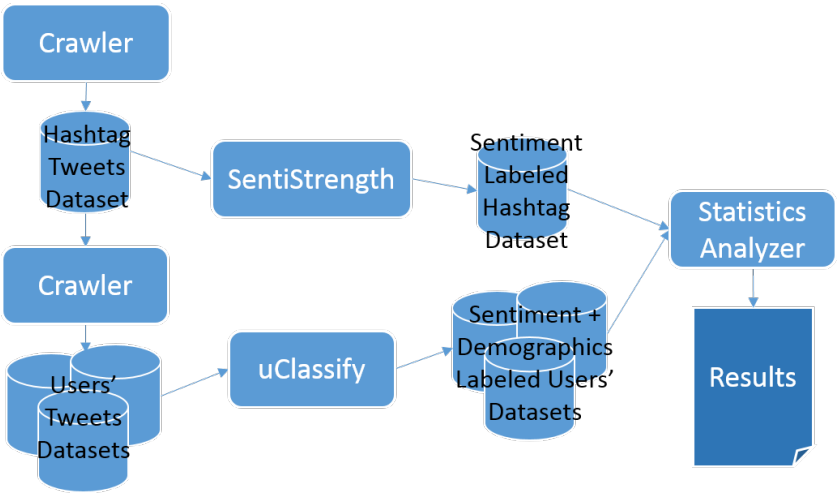

Figure 3.6: The tool for analyzing the content evolution of a stream of tweets associated with a given hashtag.

The analysis is performed by computing various statistics on the stream of tweets based on the dimensions identified in the previous subsection in the time interval of observation.

The tool is sketched in Figure 3.6 and it is constituted of four components: the **Crawler** that we have created, the open source software **uClassify**[7] and **SentiStrength**[8], and the **Statistics Analyzer** module that we have created.

As shown in Figure 3.6, an important phase is the creation of a target tweets collection that shares the considered hashtag (*Hashtag Tweets Dataset*); this is done by means of the *Crawler* over a considered time span.

Another important analysis pre-requisite is to gather a dataset for each user who has been identified as a contributor to the hashtag related tweets in the explored time span (the *Users' Tweets Datasets* in Fig. 3.6).

To the purpose of text classification we have selected the open source machine learning web service *uClassify*[7]; in our opinion this web service constitutes a good choice as it allows to easily train a custom classifier on any dataset. Based on each *Users' Tweets Dataset*, the selected text classifiers provide labels for the users on the categories that have been identified as interesting dimensions, for example: Gender, Age, Topic, etc. In Figure 3.6 the outcome of this classification phase is the named *Sentiment + Demographics Labeled Users' Datasets*.

[7]http://www.uclassify.com/
[8]http://sentistrength.wlv.ac.uk/

For the single tweet sentiment classifier a popular option is the open source software *SentiStrength*[8] that is specifically tailored to evaluate the sentiment carried by short texts. When applied to the *"Hashtag Tweets Dataset" SentiStrength* produces the *Sentiment Labeled Hashtag Dataset* as depicted in Figure 3.6.

All tweets and their associated labels constitute the input to the *Statistics Analyzer*, which has the aim of performing various statistical, demographic and sentiment analyses of the information associated with a given hashtag as it changes over the time. We have developed this software component to define some specific analysis, as it will be illustrated in Section 3.2.3.

3.2.3 Application of the Analysis Tool to a Case Study

As a demonstrative example of our tool, we have analyzed the tweets related to a solidarity movement initiated by Emma Watson. In the microblogging platform Twitter this discussion topic is identified by the hashtag #HeForShe.

We have tailored the analysis questions, presented in Section 3.2.2.1, w.r.t. the topic of this case study. In particular, we were interested to discover if the majority gender of the hashtag contributors was masculine as planned by the campaign. Another ensuing question concerned the dominant gender of the quoted sources. Moreover, we were interested to test the hypothesis if the most popular quoted sources belonged to celebrities rather than common people. The next question was related to the dominant age groups of the contributors to the promotion of gender equality. Lastly we explored the expectation that the dominant sentiment of the tweets stream was positive, based upon the noble intentions of the campaign.

3.2.3.1 Dataset Description

The tweets collection was generated using the Twitter Search API 1.1 on a weekly basis. The tweets associated with the considered hashtag, i.e., #HeForShe, were crawled, limited to those in English. Our dataset covers one month of Twitter microblogs, containing the hashtag #HeForShe, crawled from March 08 to April 08, 2015. In total it consists of 72.932 tweets,

posted by 53.700 distinct users. The retweets were included in the dataset since they comprise a Twitter mechanism to favor the propagation of information found interesting by the users; this constitutes an important information to the purpose of our analyses. Only tweets from users with protected accounts were eliminated from this dataset (about $1,5\%$ of the initial dataset). These users were deleted since their tweets are protected, thus not allowing their classification. In addition to the main dataset, we have gathered 50.445 collections of the last tweets belonging to the users who contributed to the discussion under the #HeForShe hashtag.

All tweets datasets were fed to the text classifiers of *uClassify* and to the sentiment classifier *SentiStrength* to generate labels on five categories: Gender, Age Groups, User Sentiment, Tweet Sentiment and Topic. The geographic distribution of the tweets was not considered in the analysis due to the extreme sparseness of the geographic information in the tweets and in the users' accounts.

3.2.3.2 Results of the Performed Analyses

In this section we present the results of the performed analysis with our tool.

Demographic

A first issue we addressed was to verify if the peculiarity of the hashtag (campaign) served its purpose. Although this hashtag was created for men to participate in the gender equality fight, only 18% of all tweets and only 17% of all retweets were spread by men. This gender domination trend was consistent throughout the whole time period under study at a daily granularity.

Additionally, it was interesting to compare the retweets and tweets amounts on a daily scale. Over the investigated time period, 71.829 tweets were posted on the gender equality campaign in Twitter. Out of these, 57.520 tweets were actually "retweets", and consequently only 14.309 were "original tweets" containing new content. This explains the identical oscillation pattern on a daily basis of the retweets number and the "all posts" one, observable in Figure 3.7.

Looking at the age distribution per each gender group in Figure 3.8, we found that for women the strong majority of users ($57,8\%$) was classified to the "13-17 years old" group persistently along

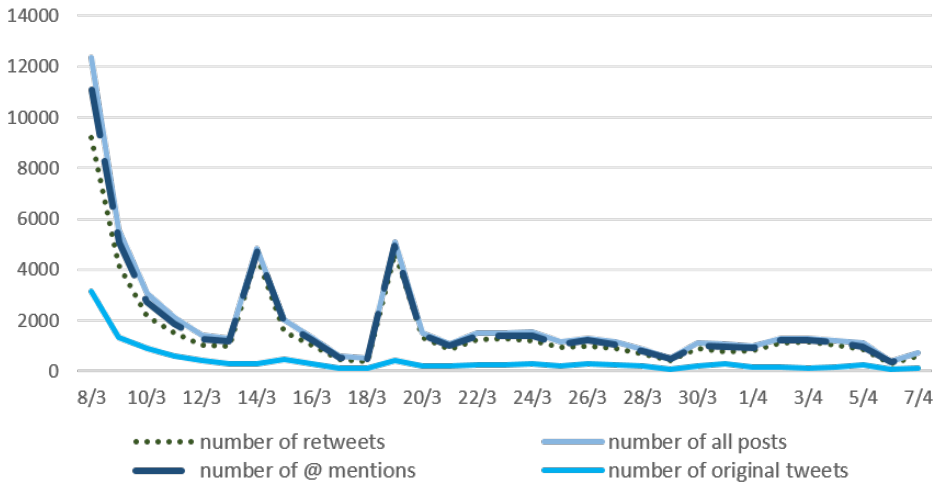

Figure 3.7: Number of all posts, original tweets, retweets and user mentions.

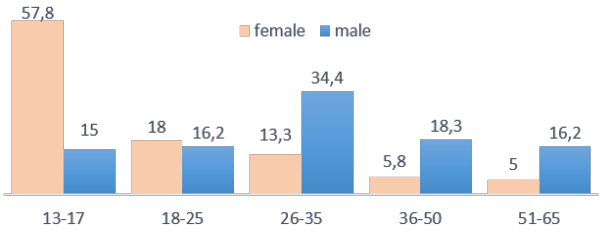

Figure 3.8: The median percentage age distribution for both genders.

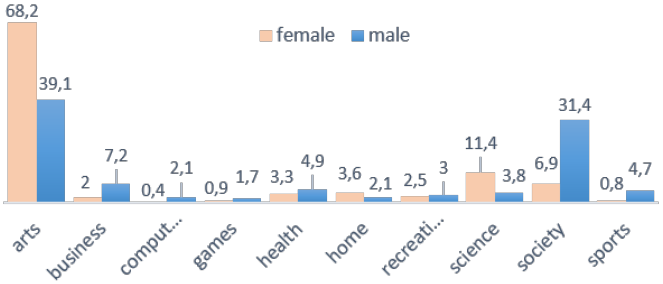

Figure 3.9: The median percentage topic distribution for both genders.

the 32 studied days. While, interestingly, a constant majority of men (34, 4%) was identified in the young adults ("26-35 years old") group, presenting an almost perfect platykurtic age distribution in the examined time span.

Sentiment

The sentiment analysis of the collected tweets and of the users who posted them showed that mainly positive people participated in this campaign, by posting mainly positive tweets. In fact, 61% of tweets were classified as positive, 7, 9% as negative and 30, 6% as neutral. By analyzing the sentiments of the users, 79% of them were found positive and 20, 8% negative. Concerning the gender ratio: 66, 5% of women and 53, 6% of men posted positive tweets.

Analysis of Topical Interests

The analysis of the users' topical interests (Figure 3.9) showed that a strong majority of users

(65, 7%) were classified as interested in the "Arts" topic. Followed by 9, 7% of users interested in "Society" and 9, 2% in "Science".

Quoted sources

Since our second exploratory study was carried out on Twitter, where quoting is performed via retweeting, we analyzed the rate at which new popular retweet sources were introduced, as well as the gender of the users who were in the top five quoted per day.

First of all, we found that for one third of the days under study two new popular retweet sources were introduced per day. Another one third of the days, only one new popularly quoted account per day was introduced. And for 28% of the days there was no new quoted sources.

Interestingly, 41, 5% of the most popular retweet sources were organizations, such as "NATO" and "United Nations". Followed by 36, 6% quoted celebrities: singers, actors, etc. News media sources were popular quoted sources in 7, 3% of cases. And non famous people in 14, 6% of cases.

As for the gender ratio of the popular retweet sources: 39% of accounts were identified as male, 26, 8% as female and 34, 2% were without gender information. 92, 8% of the Twitter account with unknown gender were organizations.

The geographical spread of the popular quoting sources was diverse. The majority, namely 34, 1% of retweet sources were from USA and 14, 6% from UK. These two locations could be a biased result due to the choice to analyze only tweets in English. 12, 2% of the retweet sources were from Asia, 9, 8% from European countries, and 4, 9% from Africa. 7, 3% were funds and organizations stating that their location is "worldwide". 17, 1% of the quoting sources had no geolocation information in their Twitter account.

Sub-topicality

To the purpose of studying the change in sub-topicality of the discussion associated with the target hashtag, first of all, we examined the top five most used hashtags jointly with #HeForShe on a daily basis. Clearly, these additional popular hashtags are related to subtopics ranging from

Figure 3.10: The #HeForShe word clouds for all the tweets for the considered time period. Left: Dataset untouched. Right: Dataset without all hashtags and user mentions.

those corresponding to the #HeForShe topic (#equalpay, #feminism, etc.) and closely related (such as International Women's Day: #iwd2015, #womensday), to those not really in the gender equality topic: #UncleOfTheYear, #facebook, #urbandictionary, etc. These additional hashtags, which are not apparently related to the topic of #HeForShe, are a demonstration of a daily shift of attention and misuse of hashtags in Twitter.

Regarding the duration of the identified hashtags popularity, only the #genderequality has been constantly used over 30 out of 32 days examined. On the second position with 16 days duration of popularity is #feminism. And on the third is the #yesallwomen with 7 days of popularity duration.

Next we looked at the words frequencies for each day under study. For the visualization we have employed the "Voyant"[9] tool, which is an open source web-based environment for texts analysis. Figure 3.10 presents the word clouds of the dataset for all 32 days studied with and without all hashtags and all user mentions, respectively. Although the additional hashtags analysis revealed general unrelated subtopics, these word clouds clearly display that the collected tweets were truly focused on promoting "men" "supporting" "women" for "equality".

3.2.3.3 Discussion of the Case Study Analysis Results

First of all, let we notice that the 80% average daily retweet rate is a positive result due to the overall goal of the campaign to create awareness and to spread information about gender equality. This goal is further supported by over 90% of all tweets including mentions of other

[9]http://voyant-tools.org/

Twitter users, which strongly helps the spread. Surprisingly, even though the campaign purpose was to activate men to fight for gender equality, the figures clearly outline that this goal was not achieved, with only 18% of tweets belonging to male users. Regarding the involved age groups, the majority of men $(34, 4\%)$ was identified in the young adults "26-35 years old" group. While for women the strong majority of users $(57, 8\%)$ were classified to the "13-17 years old" group. This is quite surprising since the topic of the campaign probably was expecting a more active participation by adult women. On the opposite side, the interest w.r.t this campaign shown by teenage girls indicates that the gender equality theme has a good level of awareness in the environments in which these girls are involved.

Even though there are plenty of emotionally colored debates on the Web around the gender equality subject, the month of tweets with #HeForShe presented a strongly positive sentiment by both female and male users. The findings of [42] were confirmed also in our study: the majority of the popular quoted Twitter users by gender are male. But the most popular quoted sources were found to be actually the Twitter accounts of organizations, funds and media sources, closely followed by celebrities. Similar to [60], we discovered that in Twitter quoting is performed without introducing any additional information such as personal opinion on the quoted subject.

3.2.4 Conclusions

In this second exploratory study on textual information in Social Media we have presented a general approach aimed to perform a qualitative and quantitative analysis of information carried by tweets gathered w.r.t a given topic. The application of this approach on the #HeForShe hashtag has proved to gather interesting and not explicitly visible insights about the spread of information on a Twitter topic, and has shown that information can follow unpredictable paths. Often the purpose of a hashtag may not be served as expected by the promoter. This study offers an interesting auxiliary tool for other fields of research, such as sociology.

3.3 Discussion

In this chapter we have presented two exploratory works on the evaluation of Social Media importance and its effect on the real world happenings. The goal of the first work was to study the correlation between the polarities of performance of companies, as expressed in Twitter posts, and the Stock Market values. The results of this study have displayed the explicitly visible correlation in the majority of cases, while the regression analysis results presented only part of the studied cases to have a statistically significant relationship. In particular, the polarity difference and the stock's closing price neither visually nor quantitatively presented a valuable relationship. Therefore, it is possible to make the conclusion that the performance polarity in tweets has less information than the general volume of tweets posted, making possibly unnecessary to undergo the time consuming and demanding task of tweets classification. Nonetheless, the fact that the volume of tweets itself has a strong correlation with Stock Market values signifies the presence of an effect of the UGC on the real world happenings.

In the second exploratory study we have proposed and developed a tool for analyzing diverse dimensions of content and characteristics of the evolution of a topic (identified by a hashtag in Twitter) in a given time interval. The results of the application of the proposed Analysis Tool on a concrete hashtag have demonstrated the expressiveness of the proposed set of dimensions and the ability of the tool to extract hidden and unexpected information in the UGC under a specific hashtag. Additionally, this study presented a strong topical diversity under a very particular hashtag.

Interestingly, both exploratory studies have identified that the UGC contains mainly positive sentiment, at least when it concerns the topics of finance and gender equality solidarity movements.

Overall, these two exploratory studies have proven the value of the UGC in the 21st century and its effect on different areas of a human life.

Chapter 4

Information Evolution Modeling

The performed exploratory studies, presented in Chapter 3, prove the importance of the information that is being spread on Social Media for different fields of study (such as, for example, finance and sociology), the presence of useful knowledge in the UGC, and the strength of the effect of the UGC on the real world happenings. Thus, these studies prove the significance of the information spread on Social Media and the hidden potential of it, making Social Media a valuable source for the study of information evolution.

Moreover, the study on the different dimensions of the information spread under a Twitter hashtag (Section 3.2), besides proving the presence of hidden knowledge in the stream of Social Media posts has emphasized the diversity of subtopics interleaved under particular hashtags in Social Media. Therefore, this leads to a conclusion that a hashtag itself cannot be assumed to represent an ememe. This inference has highlighted the necessity for an approach for modeling the spread of textual information in Social Media streams for the objective of identifying and extracting the core units of information from these streams, aka ememes, for the purpose of tracking and analyzing the information evolution; this served as the motivation for the main contribution of the current thesis work.

Thus, the core work of this PhD thesis is three-fold: 1) knowledge representation part: Section 4.1 describes the proposed approach for modeling textual information in a Social Media stream; 2) knowledge extraction part: Section 4.2 presents the proposed approach for the identification

of the ememes in Social Media streams; and finally, 3) information evolution tracking and analyzes part: Chapter 5 describes the two sets of measures proposed for the task of the quantitative and qualitative evaluation of the evolution of information in time.

In this chapter, the proposed process for modeling Social Media streams (Section 4.1) and identifying ememes (Section 4.2) is defined in-depth and formally. Meanwhile, a general overview of the overall process is the following: given a finite stream of Social Media posts collected over a finite temporal window we define an iterative procedure to construct a graph (Graph-of-Words) that formally represents this stream. To this aim, we first define a graph associated with the first post in the stream, then we iteratively update it by generating and adding (one by one) the graphs of the next posts of the considered stream. We apply then to the obtained graph the k-core degeneracy algorithm to extract the core of the graph that we assume to represent an ememe or a set of ememes. Then, we repeat this procedure for the subsequent stream of Social Media messages of equal temporal size and we obtain its k-core. This procedure is further repeated for a pre-defined number of streams of Social Media posts collected in temporally equal and sequential time windows. It is important to notice that each of the temporally equal Social Media streams may contain a different number of posts and a different number of ememes.

4.1 Knowledge Representation

In the following the issue of formally representing a textual information stream is addressed. An extensive review of the prior works and the variety of approaches to the task of textual information representation in the context of content-centered Information Propagation is presented in Section 2.2.2. While the most popular model for text representation remains the bag-of-words one, the graph-based models encompass more information about the relationship between the nodes and the edges. Thus, in the core work of the present PhD thesis for the task of modeling Social Media streams we employ a graph-based approach.

In Section 4.1.1 we present a brief review of the prior works on the adoption of the graph-based representation of texts in the Social Media related fields. Further on, in Section 4.1.2 we present

the formal definition of the Graph-of-Words approach, which we employ in the core work of this thesis for the task of modeling Social Media streams. While in Section 4.1.3 we present the the formal in-depth description of the full process of modeling a Social Media stream with the Graph-of-Words approach.

4.1.1 Graph-Based Representation of Textual Information in Social Media

In the following we present some research works performed in the context of Social Media. In particular, we critically review some important works that employ graph-based representations of textual information in the fields of Topic/Event Detection. Although the focus of these studies differs from the one of the core work of this thesis (we reason about this matter further on) these works serve as good basis for the evaluation of the performance of the graph-based representation of information in the context of Social Media.

A graph-based representation of micro-blogging messages and documents has been previously used for the task of Event Detection. In [11] the community detection methods are used on a KeyGraph (proposed by [64]) for the detection of events in documents. Similarly, in [10] the k-core degeneracy of a Graph-of-Words has been applied for the task of identifying sub-events in Twitter. Both the KeyGraph and the Graph-of-Words are built with nodes representing terms and edges representing the co-occurrences of these terms in the same document/tweet. However, there is a slight difference in the construction techniques of these two graph types. In the KeyGraph the nodes represent the keywords in a document that have a high document frequency, and the edges between these nodes are constructed if three conditions are met: (i) the number of co-occurrences of the two keywords in a document has to be above a threshold, (ii) the probability of seeing the first keyword in the document if the second keyword exists in the document has to be above a threshold, and (iii) the inverse probability of the second condition has to be above a threshold. In the Graph-of-Words, instead, the nodes are all unique words of a tweet and the edges are built for all the co-occurrences of these unique words in the same tweet. Additionally, there is a difference between these two approaches in terms of

the techniques used for the event detection and description: in the KeyGraph the community detection methods are used analogously to their usage for the Social Network analysis, and in the Graph-of-Words approach the technique named k-core graph degeneracy is applied.

We find important to stress that even though events and (e)memes are similarly defined as a set of interdependent terms, the main difference is that an event is constrained by time while the notion of a meme is not. On contrary, one of the main characteristics of a meme is it's persistence in time. For this main reason, the measures of an ememe that we propose in the core work of this thesis are the means to verify whether a message gone viral is a true ememe. While the same measures might not be useful for different tasks such as Event Detection. Nonetheless, an event may give rise to a number of (e)memes related to the event itself or to some aspects of it but this is not mandatory.

Topic Detection, which is a contiguous research field to Event Detection, has also been approached with graph-based techniques. For example, in [9] a KeyGraph approach is used to build a keywords co-occurrence graph, on which the community detection algorithm is used for the task of topic features detection. Then the identified topic features are used to assign documents to topics.

Another application of the KeyGraph for Topic Detection and Tracking is presented in [7] for a study in the domain of the news Web pages. Web pages are clustered based on their timestamps. For each temporal cluster a KeyGraph is constructed, where the nodes are frequent words and edges represent the high co-occurrence of these words in one sentence. Each maximally connected subgraph is called a basic concept. The basic concepts from the KeyGraph allow to identify sub-clusters which present sub-topics in the temporal cluster of web pages. The basic concepts are measured by cosine similarity inside each cluster and with the basic concepts of the sequential temporal clusters, thus, allowing for topics to be tracked.

A graph-based approach for analyzing the dynamic changes of the online debates evolving over time was proposed in [8]. A graph is built to reflect the relationships between terms found in the texts of the online debate. It is composed of the links between all possible combinations of the term pairs within a sentence. Then blocks of co-occurring terms within the built graph

are found with the depth-first search tree. These blocks are assumed to represent the subjects discussed within a broad topic of the online debate. Four graph property measures (diameter, effective diameter, average vertex degree and the Densification Power Law plot) from [65] are used to see how the graph is evolving over time. This work proved that the graph structures can be used to observe the successive appearance and demarcation of subjects along time.

A combination of terms' life cycle and the social relationships in the network is used in [6] to create a topic graph of co-occurring terms in order to obtain a set of emerging topics. First of all, Twitter posts were formalized as vectors of terms with their relative frequencies. Then the Page Rank algorithm was used for the analysis of the social relationships in the directed graph of active users with the purpose of identification of the authority of the users. Aging theory based on frequency in tweets and user authority is used to model the life-cycle of each term. A set of emerging terms is selected by ranking the keywords based on their life status with either proposed supervised or unsupervised techniques. And as the final step, a topic graph is created that links the extracted emerging terms with their relative co-occurring terms for the identification of the emerging topics. In [6], a topic is defined as a minimum set of terms semantically related to an emerging keyword(s) within a defined time interval.

In contrast to (e)memes, the notion of a topic has a much higher generality. A topic may implicate a number of ememes at different points in time.

The aforementioned works are related to the fields that are similar to the one approached in the current thesis, and, therefore, they serve as evidence of the success of the graph-based representations of textual information in the Social Media environment.

4.1.2 Graph-of-Words

As the first part of the core work of this PhD thesis, to model a stream of Social Media posts we propose to employ a graph-based representation called Graph-of-Words, the unweighted and directed version of which was first introduced in [66]. In our approach we employ the weighted and undirected Graph-of-Words, as reported in Definition 1.

Definition 1 (Graph-of-Words). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a weighted undirected graph, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges. We denote by $\mathcal{W}_v : \mathcal{V} \rightarrow \mathbb{R}^+$ the weight mapping on the nodes, and by $\mathcal{W}_e : \mathcal{E} \rightarrow \mathbb{R}^+$ the weight mapping on the edges. The nodes of the graph represent unique terms, and edges connect terms that co-occur in the considered text.*

In a weighted graph the *weighted* degree of a node v is computed as the sum of the weights of its incident edges [67]. We assume that function $\mathcal{W}_v$ computes the weighted degree of a node, i.e. a node's weight is its weighted degree.

4.1.3 Representation of a Stream of Social Media Posts

In this section we present a detailed and formal description of the proposed process for modeling a stream of Social Media posts.

We assume to have a finite stream of Social Media posts. We partition this finite stream of Social Media posts into N sequential substreams of posts appearing in *time windows* of equal duration. We denote these sequential substreams as TS_1, TS_2, ..., TS_n, ..., TS_N. The granularity of the considered time window may depend on both the topic and the purpose of the analysis, and it could be an hour, a day, a week, etc. The smaller the granularity the more potential ememes can be identified, but the noise can also increase. The larger the granularity the simpler it is to verify the persistence in time of the identified ememes. In Chapter 6 we will demonstrate how different time granularities impact the quality, quantity and topical aspects of the identified ememes. Introducing a partitioning of an observed stream of posts into substreams TSs is necessary in order to perform an analysis of the evolution of ememes over time.

In the following we explain the process of constructing the weighted undirected Graph-of-Words G_{TS_n} (see Definition 1 in Section 4.1.2) representing the TS_n substream of posts associated with the time window n; the overall process is illustrated in Figure 4.1. We denote by G_{TS_n} the Graph-of-Words representing the target substream of posts TS_n; we denote by $g_{p_i^{TS_n}}$ the Graph-of-Words representing each post of the TS_n. Each post $p_i^{TS_n}$ of the target TS_n is preprocessed (see Section 6.2 for details), including its conversion to a bag-of-words of single terms of the

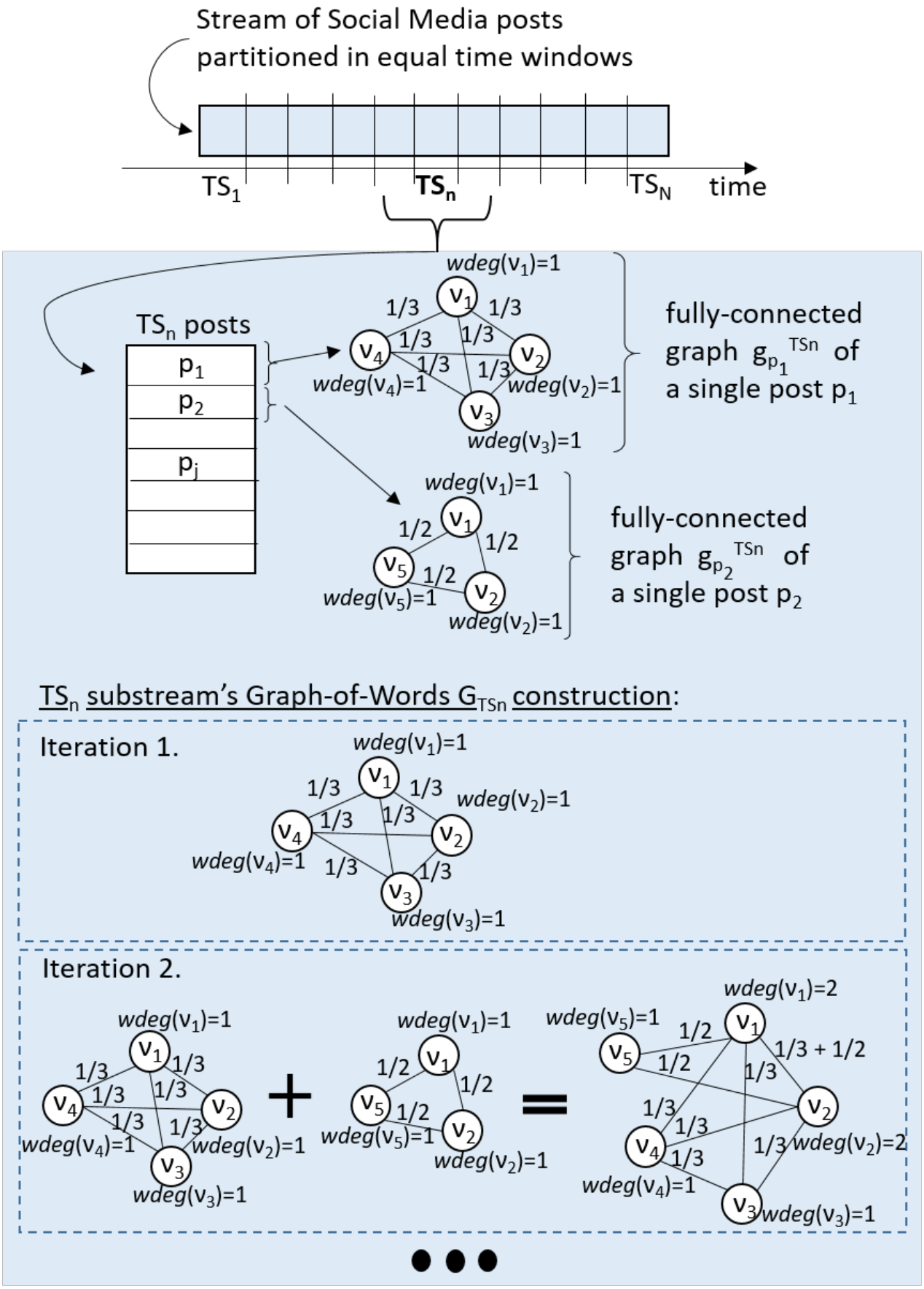

Figure 4.1: General scheme of the process of generating the graph-based representation of the TS_n substream of Social Media posts.

post. Then we create the G_{TS_n} of the target TS_n by means of an iterative process. We start by representing the first preprocessed post $p_1^{TS_n}$ of the target substream TS_n as a fully-connected graph $g_{p_1^{TS_n}}$, where the nodes represent the words of the post. To treat all words in the single post equally we set all the weighted degrees in $g_{p_1^{TS_n}}$ to 1. As outlined in Section 4.1.2, the weighted degree of a node is defined as the sum of the weights of its incident edges. Therefore, to share the weighted degree of each node over its incident edges, we compute the weight of each edge as $w = \frac{1}{l-1}$, where l is the number of nodes in the fully-connected $g_{p_1^{TS_n}}$. The obtained Graph-of-Words $g_{p_1^{TS_n}}$ is the initial G_{TS_n}. Then we repeat the same process for the next post $p_2^{TS_n}$in the target substream TS_n and obtain its Graph-of-Words representation $g_{p_2^{TS_n}}$. The initial G_{TS_n} is then merged with $g_{p_2^{TS_n}}$: the weights of all the edges of the existing nodes in G_{TS_n} are incremented by the weights of the edges in the second post's graph $g_{p_2^{TS_n}}$. For those nodes that did not exist in the initial G_{TS_n} new nodes and corresponding edges are created by maintaining the edges' weights from the $g_{p_2^{TS_n}}$. The weighted degrees of the nodes are recalculated by summing up all the weights on the edges incident to each node. This procedure is iterated for every post in the TS_n to obtain a final Graph-of-Words G_{TS_n} of the Social Media posts substream TS_n:

$$G_{TS_n} = g_{p_1^{TS_n}} + g_{p_2^{TS_n}} + \ ... \ + g_{p_i^{TS_n}} + \ ... \ + g_{p_M^{TS_n}},$$

where M is the number of Social Media posts in the substream TS_n.

Based on the process illustrated above, the larger and the more general (in terms of topics) is the Social Media posts substream TS_n, the bigger is the Graph-of-Words associated with it.

We normalize the weighted degrees of the nodes of the final Graph-of-Words G_{TS_n} for a given TS_n in the $[0, 1]$ interval by dividing all the nodes' weights by the largest node weight in the G_{TS_n}. We normalize the weights on the edges similarly by the largest edge weight in the G_{TS_n}.

The described process is then repeated for all the subsequent TSs in the given stream of Social Media posts. Thus, we obtain a temporal sequence of Graph-of-Words representations of all substreams:

$$TS_1 \rightarrow G_{TS_1}, ..., TS_n \rightarrow G_{TS_n}, ..., TS_N \rightarrow G_{TS_N}.$$

4.2 Knowledge Extraction

Once the sequential substreams of Social Media posts have been represented as Graphs-of-Words, the next step of the core work of this thesis is to extract for each substream the core units of information - the ememes of that substream. For this purpose we propose to use the k-core graph degeneracy technique, formally defined in Section 4.2.1. Furthermore, Section 4.2.2 formally presents the full process of identifying the ememes basis with the k-core graph degeneracy, and Section 4.2.3 presents the two proposed formal definitions and representations of an ememe - the crisp and the fuzzy ones.

4.2.1 *k*-Core Graph Degeneracy

In Definition 2, we introduce the concept of k-core graph degeneracy [68], which we employ to identify the potential ememes in the Graph-of-Words representing the stream of Social Media posts.

Definition 2 (***k***-core)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and k be a positive real number. A subgraph $\mathcal{H}_k = (\mathcal{V}', \mathcal{E}')$ of the graph $\mathcal{G}$, induced by the subset of nodes $\mathcal{V}' \subseteq \mathcal{V}$ and the subset of edges $\mathcal{E}' \subseteq \mathcal{E}$, is called a k-core or a core of order k of graph $\mathcal{G}$ iff $deg(v) \geq k$, $\forall\ v \in \mathcal{V}'$. And $\mathcal{H}_k$ is the maximum subgraph which satisfies the property $deg(v) \geq k$, $\forall\ v \in \mathcal{V}'$, i.e. it cannot be augmented without losing this property.*

More specifically, in our work we define and apply an extension of the k-core degeneracy algorithm, which considers the weighted degrees of nodes in the Graph-of-Words instead of the classic degrees of nodes. Thus, the k-core of a graph is the maximal connected subgraph where each node has a weighted degree of at least k within the subgraph. The set of all k-cores of the graph forms the *k-core decomposition* of a graph.

The k-degenerated graph notion was introduced by Stephen Seidman in 1983 [68], and later, in 2003, a work in [69] has introduced an efficient algorithm for determining the core decompositions of a graph with $\mathcal{O}(max(m, n))$ time complexity, which equals $\mathcal{O}(m)$ for connected networks, where n is the number of nodes and m is the number of edges. This algorithm is based on the recursive elimination of all nodes and edges incident with them of degree less than k.

4.2.2 Ememes Basis Identification

As the result of the process proposed for the task of modeling Social Media streams, described in Section 4.1.3, we have obtained a temporal sequence of Graph-of-Words representations of all substreams:

$$TS_1 \rightarrow G_{TS_1}, ..., TS_n \rightarrow G_{TS_n}, ..., TS_N \rightarrow G_{TS_N}.$$

Now we apply the k-core degeneracy algorithm (see Definition 2) to each final and normalized graph G_{TS_n}: in Section 6.3 a discussion related to the setting of parameter k is presented. The application of the k-core degeneracy process is aimed at identifying the core vocabulary of the ememes in the TS_n. In fact, the outcome of this algorithm is constituted by a set of subgraphs: the nodes of each subgraph represent a set of words that co-appear predominantly in the observed Social Media posts substream. We claim that these subgraphs may be considered as a basis of the representations of ememes of the substream TS_n.

Definition 4.1 (***k***-core of the Graph-of-Words as an Ememes Basis)**.** *An ememes basis B^{TS_n} is the set of subgraphs (one or more) generated by applying the k-core graph degeneracy to the final Graph-of-Words G_{TS_n} associated with the substream TS_n.*

Thus, the *ememes basis* B^{TS_n} represents the core granules of information in the stream TS_n. To further verify the quality of the obtained subgraphs in the *ememe basis* B^{TS_n} an outliers detection technique is applied to identify nodes with significantly higher weighted degree, which we call "central concepts". The extraction of "central concept(s)" forces the subgraph in which

they were identified to fall into a number of additional subgraphs. This way we can identify the ememes that spin around the same central concepts or entities.

4.2.3 Formal Representation of an Ememe

As outlined in the previous section, an ememes basis B^{TS_n}, generated by the application of the k-core graph degeneracy algorithm to the Graph-of-Words G_{TS_n} associated with a substream TS_n, constitutes the basis for the formal representation of the ememes of TS_n. We propose then to derive from an ememes basis B^{TS_n} a formal representation of the associated ememes. In particular, we propose two representations of an ememe: a crisp one, as formally described in Definition 4.2, and a fuzzy one, as formally described in Definition 4.3.

4.2.3.1 Crisp Representation of an Ememe

In the following Definition 4.2 we present the formal representation of an ememe as a crisp subset of terms.

Definition 4.2 (Ememe as a Crisp Subset)**.** *Given a Social Media posts substream TS_n, an ememe $m_i^{TS_n}$ denotes the crisp subset of terms of the ith subgraph from the ememes basis B^{TS_n}.*

Thus, the crisp representation of an ememe is basically the bag-of-words, derived from the nodes representing terms in the target ith subgraph of the ememes basis B^{TS_n} (Definition 4.1).

4.2.3.2 Fuzzy Representation of an Ememe

Definition 4.3 presents the formal representation of an ememe as a fuzzy subset of terms.

Definition 4.3 (Ememe as a Fuzzy Subset)**.** *Given a Social Media posts substream TS_n, an ememe $m_i^{TS_n}$ denotes the fuzzy subset of terms of the ith subgraph from the ememes basis B^{TS_n}.*

$V_{m_i^{TS_n}}$ denotes the vocabulary employed in the ememe representation and is composed of the terms associated with the nodes of the ith subgraph of the ememes basis B^{TS_n} (Definition 4.1).

The degree of membership of each word is the normalized weight of the corresponding node in the ith subgraph of the ememes basis B^{TS_n}.

$$\mu_{m_i^{TS_n}}\colon V_{m_i^{TS_n}} \rightarrow [0,1],$$

where $\mu_{m_i^{TS_n}}$ denotes the membership function of the ememe $m_i^{TS_n}$.

The fuzzy subset representing the ememe $m_i^{TS_n}$ is denoted as follows:

$$F_{m_i^{TS_n}} = \{\mu_{m_i^{TS_n}}(v_1)/v_1,\, \mu_{m_i^{TS_n}}(v_2)/v_2,\, ...\, \mu_{m_i^{TS_n}}(v_j)/v_j\},$$

where the notation $\mu_{m_i^{TS_n}}(v_j)/v_j$ denotes a term $v_j \in V_{m_i^{TS_n}}$ and its associated membership value.

Furthermore, we introduce a fuzzy relation $R_{m_i^{TS_n}}$ associated with ememe $m_i^{TS_n}$; the membership function of the fuzzy relation is defined on the cartesian product $V_{m_i^{TS_n}} \times V_{m_i^{TS_n}}$ of the vertices of the ith subgraph from the ememes basis B^{TS_n} (Definition 4.1). The membership value of a pair of vertices (v_i,v_j), is the weight of the edges connecting the two nodes in the considered subgraph.

4.3 Discussion

In this section we have presented the proposed approach to modeling textual information streams in Social Media, and to identifying and extracting the core units of information, *ememes*, from these Social Media streams.

In the first part of this chapter, before presenting the graph-based approach proposed to use in this thesis, we have critically reviewed a number of prior works employing graph-based representations of Social Media posts in related fields such as Topic/Event Detection. Besides the conclusion from these reviewed studies that graph-based text representation is a successful technique in the context of the Social Media environment, another importance inference emerged -

the novelty of the current study. Although we discuss in the Section 4.1.1 the major differences between these works in relation to our, nonetheless, the major novelty of this PhD thesis work is due to the proposed conceptual definition of an ememe (two definitions - crisp and fuzzy) and due to the proposed sets of measures for the evaluation of the evolution of information, which we present in the following Chapter 5. Nonetheless, to our knowledge this is the first work to propose to use Graph-of-Words approach with the subsequent extended k-core graph degeneracy for the task of identification and extraction of ememes from Social Media posts streams.

The purpose of the proposal of two distinct conceptual definitions of an ememe, the crisp and the fuzzy ones, is due to the following number of reasons. First of all, by virtue of the membership values of each term, the fuzzy ememe definition allows to understand the comparative value of each term in the ememe, while in the crisp ememe definition each term has an equal value in the ememe. From this same characteristic of the fuzzy ememe follows its another advantage - the membership value of each term in the fuzzy-defined ememe allows a more detailed evaluation of the evolution of ememes over time by allowing to evaluate not only the difference in the present terms in the ememes of different substreams but additionally to evaluate the shift in the importance of the ememes' terms in two different substreams. This matter will be discussed in more details in the next chapter. Nonetheless, the main drawback of the fuzzy definition of an ememe is due to the additional complexity introduced with the definition of the membership values for each term in the ememe. Thus, the crisp definition of an ememe allows most of the same evaluations of the information evolution as the fuzzy definition with lower complexity at the cost of lower precision and lower detailing of the analysis. Depending on the application and the purpose of the information evolution study, a fuzzy ememe definition can bring additional and important knowledge.

Chapter 5

Information Evolution Tracking

Once all of the sequential substreams of Social Media posts have been represented with a Graph-of-Words approach (Section 4.1) and the enhanced k-core degeneracy algorithm has been applied to identify the *ememes* in each substream of Social Media posts (Section 4.2), as the third part of the core work of this PhD thesis we propose two sets of measures - crisp set based (Section 5.1) and fuzzy subset based (Section 5.2) - for the tasks of tracking and analyzing the evolution of information in Social Media streams.

The proposed measures allow: a) to validate whether the identified ememe possesses the main characteristic of a meme, i.e. persistence in time; b) to track and measure the change of ememes as they propagate through time; c) to identify the most dominant ememe(s) for each time window in the target Social Media posts stream; d) to evaluate how well is the ememe represented in its substream of Social Media posts; e) to understand the spreading power of the instant sharing functionalities of the Social Media platforms.

To the purpose of formally defining the proposed measures, both fuzzy and crisp, we formally define a Social Media post.

Definition 5.1 (Representation of a Social Media post)**.** *A Social Media post* $p_j^{TS_n}$ *is a set of words.*

Each Social Media post $p_j^{TS_n}$ of the TS_n is represented as the set of words that this post

contains, excluding duplicate words and stopwords (see Section 6.2 for details). In this case the membership function is binary: the words of the post are either included in the set or not. Due to the fact that Social Media posts are usually very short, the crucial point is the presence of a word and not the count of its occurrences.

We emphasize the fact that all the presented definitions of crisp and fuzzy measures are applicable to any Social Media platform.

5.1 Crisp Measures

In the following we present a set of crisp measures for the crisp representation of ememes, in accordance with Definition 4.2.

5.1.1 Crisp Relatedness

To identify how well an ememe $m_i^{TS_n}$ is represented in each post $p_j^{TS_n}$ in the substream of Social Media posts TS_n , we propose a measure called *crisp relatedness*.

Definition 5.2 (Crisp Relatedness of a crisp-defined ememe to a Social Media post)**.** *The relatedness of a crisp-defined ememe $m_i^{TS_n}$ to a Social Media post $p_j^{TS_n}$ is the degree of inclusion of the set representing the ememe $m_i^{TS_n}$ to the set representing the post $p_j^{TS_n}$.*

$$CrispRelatedness(m_i^{TS_n}, p_j^{TS_n}) = \frac{|m_i^{TS_n} \cap p_j^{TS_n}|}{|p_j^{TS_n}|},$$

Since a Social Media post $p_j^{TS_n}$, likewise a crisp-defined ememe $m_i^{TS_n}$, is formally represented as the classic set of unique terms that compose it, the *crisp relatedness* score is computed as the count of the terms in common between the ememe $m_i^{TS_n}$ and the post $p_j^{TS_n}$ divided by the number of terms in the post $p_j^{TS_n}$.

5.1.2 Crisp Coverage

To identify the proportion of Social Media posts $p_j^{TS_n}$ that contain an ememe $m_i^{TS_n}$ in a substream TS_n, we propose the measure called *crisp coverage*.

Definition 5.3 (Crisp Coverage)**.** *The crisp coverage of a crisp-defined ememe $m_i^{TS_n}$ is the percentage of the TS_n posts p^{TS_n} that have a crisp relatedness score to the ememe $m_i^{TS_n}$ above a threshold J.*

The threshold is to be set in accordance to the analyzed Social Media stream characteristics, such as size and topic aspects. In Section 6.4.1.2 we present a discussion regarding the *crisp coverage* threshold selection for the concrete examples of datasets.

A relevant aspect of the *crisp coverage* measure is that a Social Media post $p_j^{TS_n}$ may include more than one ememe of its substream TS_n.

5.1.3 Crisp Prevalence

To evaluate the importance and value of each crisp-defined ememe $m_i^{TS_n}$ in comparison to the other ememes in the same substream TS_n, we propose a measure called *crisp prevalence*. This measure enables the identification of the most dominant ememe for each studied substream.

Definition 5.4 (Crisp Prevalence)**.** *Each Social Media post $p_j^{TS_n}$ from TS_n is evaluated to establish to which crisp-defined ememe it has the highest score of crisp relatedness. Then the crisp prevalence of an ememe in a TS_n is calculated by computing the percentage of posts from TS_n that have the highest crisp relatedness score to that ememe.*

In simpler words, the calculation of the *crisp prevalence* for each ememe of the substream is performed with the following process. Firstly, we assign to each post $p_j^{TS_n}$ the ememe $m_i^{TS_n}$ that has the highest *crisp relatedness* score to it. Subsequently, the *crisp prevalence* measure for $m_i^{TS_n}$ is calculated as the number of posts in TS_n assigned to $m_i^{TS_n}$ divided by the total number of posts in TS_n.

5.1.4 Crisp Fidelity

To verify the persistence in time of an ememe, we propose a measure called *crisp fidelity*, presented in Definition 5.5.

Definition 5.5 (Crisp Fidelity)**.** *Crisp fidelity is the value of similarity between the crisp-defined ememe* $m_i^{TS_n}$ *and the crisp-defined ememe* $m_j^{TS_{n+m}}$*.*

This measure is calculated as the Szymkiewicz-Simpson coefficient, or overlap coefficient, which is a similarity measure between the crisp set of words composing an ememe $m_i^{TS_n}$ from a substream TS_n and the crisp set of words composing an ememe $m_j^{TS_{n+m}}$ from another subsequent or preceding substream TS_{n+m}.

$$CrispFidelity(m_i^{TS_n}, m_j^{TS_{n+m}}) = \frac{|m_i^{TS_n} \cap m_j^{TS_{n+m}}|}{min(|m_i^{TS_n}|, |m_j^{TS_{n+m}}|)},$$

5.1.5 Crisp Virality

With the fast life pace of the XXI century, people tend to share content through Social Media instant "sharing" functionalities, such as "retweets" in Twitter, "share now" button in Facebook, etc. This allows with one click to spread an interesting or important piece of information. People often prefer this type of information sharing approach rather than rewriting information themselves, even if they agree (or find interesting) only part of the shared content.

Thus, we propose a measure called *crisp virality* to understand the viral information diffusion power of the "share" functionality in Social Media; and whether an ememe can be viral without having a high instant sharing rate. Another interesting aspect that the *crisp virality* measure can bring insight upon is the ratio of the ememe extracted from the original posts in the ememe generated from reposts.

Definition 5.6 (Crisp Virality)**.** *Crisp virality is the value of the similarity between crisp-defined ememe* $m_i^{TS_n}$ *extracted from the original posts of* TS_n *and crisp-defined ememe* $rm_j^{TS_{n+m}}$ *extracted from the reposts of* TS_n*.*

$$m_i^{TS_n} \sim rm_j^{TS_n} \in [0, 1]$$

To compute the *crisp virality* measure, first of all we create a separate Graph-of-Words for only the original Social Media posts of the substream TS_n, excluding the reposts, and a separate Graph-of-Words for the reposts of the substream TS_n (see Section 4.1.3 for the full process of a Graph-of-Words representation of a substream TS_n). Then we perform the enhanced k-core graph degeneracy (as described in Section 4.2.2) on both graphs separately and identify the crisp-defined ememes in the original posts and in the reposts. Finally, we calculate the *crisp fidelity* measure between each crisp-defined ememe $m_i^{TS_n}$ extracted from the original posts of substream TS_n and each crisp-defined ememe $rm_j^{TS_n}$ from the reposts of the same substream TS_n, to understand whether or not the reposts spread the same ememes as the original posts.

5.1.6 Crisp Longevity

To understand the length of the existence of a crisp-defined ememe $m_i^{TS_n}$, we propose a measure called *crisp longevity*.

Definition 5.7 (Crisp Longevity)**.** *Crisp longevity is the length of a lifespan of a crisp-defined ememe* $m_i^{TS_n}$*.*

Crisp longevity is computed as the number of TSs, out of the overall N number of TSs studied, in which an ememe $m_i^{TS_n}$ was preserved up to a given threshold of *crisp fidelity*. In Section 6.4.1.2 we present a discussion regarding the *crisp longevity* threshold selection for the concrete examples of datasets.

5.2 Fuzzy Measures

In this section we define a set of fuzzy measures that we propose to characterize, track in time and analyze an ememe represented as a fuzzy subset of terms (Definition 4.3).

5.2.1 Fuzzy Relatedness

Similarly to the *crisp relatedness* measure presented in Definition 5.2 (Section 5.1.1), in Definition 5.8 we formally present a *fuzzy relatedness* measure for the identification of how well a fuzzy-defined ememe $m_i^{TS_n}$ is represented in each post $p_j^{TS_n}$ in the substream of Social Media posts TS_n.

Definition 5.8 (Fuzzy Relatedness of a fuzzy-defined ememe to a Social Media post)**.** *The relatedness of a fuzzy-defined ememe $m_i^{TS_n}$ to a Social Media post $p_j^{TS_n}$ is the degree of inclusion of the fuzzy subset representing the ememe $m_i^{TS_n}$ to the crisp set representing the Social Media post $p_j^{TS_n}$.*

$$FuzzyRelatedness(m_i^{TS_n} \subseteq p_j^{TS_n}) = \frac{\sum Count(p_j^{TS_n} \cap m_i^{TS_n})}{\sum Count(p_j^{TS_n})} = \frac{\sum\limits_{t \in V_{p_j^{TS_n}}} min(\mu_{p_j^{TS_n}}, \mu_{m_i^{TS_n}})}{\sum\limits_{t \in V_{p_j^{TS_n}}} \mu_{p_j^{TS_n}}},$$

where t is the term in the $p_j^{TS_n}$ post's vocabulary $V_{p_j^{TS_n}}$.

The *fuzzy relatedness* of a fuzzy-defined ememe to a Social Media post is calculated based on the number of terms in common between the ememe's fuzzy subset and the post's classical set. The difference between the *fuzzy relatedness* and the *crisp relatedness* calculations is due to the difference in the two representations of an ememe. Since the fuzzy-defined ememe's terms are weighted (the degree of membership of these terms to the ememe fuzzy subset), the *fuzzy relatedness* score is calculated as the sum of the weights of the terms that the post has in common with the ememe, divided over the number of terms in the post.

5.2.2 Fuzzy Coverage

Similarly to the *crisp coverage* (Definition 5.3, Section 5.1.2), the measure of *fuzzy coverage*, formally presented in the Definition 5.9, is used to identify the proportion of Social Media posts $p_j^{TS_n}$ in the TS_n substream that "include" the fuzzy-defined ememe $m_i^{TS_n}$.

Definition 5.9 (Fuzzy Coverage)**.** *The fuzzy coverage of a fuzzy-defined ememe* $m_i^{TS_n}$ *is the percentage of the* TS_n *posts* p^{TS_n} *that have a relatedness score to the ememe* $m_i^{TS_n}$ *above a threshold J.*

In Section 6.4.2.2 we present a discussion related to the setting of the threshold J for the *fuzzy coverage*. This threshold setting depends on the different topics under study and the different purposes of the analysis to be performed.

This measure allows to assume that a Social Media post may include more than one ememe.

The difference between the *fuzzy coverage* in relation to the *crisp coverage* is due to the different calculations of the *fuzzy/crisp relatedness* measures.

5.2.3 Fuzzy Prevalence

In Definition 5.10, similarly to the *crisp prevalence* (Definition 5.4, Section 5.1.3) we propose a measure of *fuzzy prevalence* to evaluate the importance of each ememe in comparison to other ememes of the target TS_n. *Fuzzy prevalence* can be used to identify the most dominant ememe for each Social Media posts substream.

Definition 5.10 (Fuzzy Prevalence)**.** *Each Social Media post* $p_j^{TS_n}$ *from* TS_n *is evaluated to establish to which ememe it has the highest score of fuzzy relatedness. Then the fuzzy prevalence of an ememe in a* TS_n *is calculated by computing the percentage of posts from* TS_n *that have the highest fuzzy relatedness score to that ememe.*

The measure of ememe's *fuzzy prevalence* allows to comparatively evaluate the predominance of each fuzzy-defined ememe identified in the studied TS_n with respect to the other fuzzy-defined ememes of that TS_n.

This measure for all ememes of the TS_n in the best case scenario adds up to a 100%.

The measure of the fuzzy-defined ememe's *fuzzy prevalence* is calculated by choosing for each Social Media post the ememe to which it has the highest fuzzy relatedness score, in other words,

which ememe is the most similar to this post. Hence, we obtain the importance of each ememe for the target TS_n through the percentage of the TS_n posts that each ememe "represents". And by comparing all TS_n ememes' *prevalence* values we identify the most dominant ememe of the TS_n.

Once again, the difference between the calculation of the *fuzzy prevalence* in comparison to the *crisp prevalence* is due to the different calculation of the *fuzzy/crisp relatedness*.

5.2.4 Fuzzy Fidelity

In Definition 5.11, similarly to the *crisp fidelity* (Definition 5.5, Section 5.1.4), we introduce a measure called *fuzzy fidelity* for the evaluation of the similarity between two fuzzy-defined ememes. *Fuzzy fidelity* is a comparative measure. It is calculated for an ememe of interest from one substream in relation to all other ememes in previous and subsequent substreams. The procedure of calculating the *fuzzy fidelity* measure will be presented on two hypothetic fuzzy-defined ememes $m_i^{TS_n}$ from the substream TS_n and $m_j^{TS_{n+m}}$ from one of the subsequent substreams TS_{n+m}.

Definition 5.11 (Fuzzy Fidelity)**.** *Fuzzy fidelity is the value of similarity between the fuzzy-defined ememe* $m_i^{TS_n}$ *and the fuzzy-defined ememe* $m_j^{TS_{n+m}}$*.*

The similarity between the fuzzy subsets representing the two ememes is computed as suggested in [70]:

$$S(m_i^{TS_n}, m_j^{TS_{n+m}}) = 1/h \sum_{k=1}^{h} \left(\frac{\min(\mu_{m_i^{TS_n}}(v_k), \mu_{m_j^{TS_{n+m}}}(v_k))}{\max(\mu_{m_i^{TS_n}}(v_k), \mu_{m_j^{TS_{n+m}}}(v_k))}\right),$$

where v_k is a term of the vocabulary of the ememe, and h is the number of terms in the vocabulary of the ememe with the highest number of terms.

The measure of *fuzzy fidelity* allows to verify the ememe's main characteristic: its persistence in time. It allows to identify whether the chosen fuzzy-defined ememe of interest is preserved in the previous and next time windows.

5.3 Discussion

In this section we have presented the proposed sets of crisp and fuzzy measures for the third and final part of the core work of the current PhD thesis - the task of quantitative and qualitative evaluation of the evolution of information on Social Media. These measures, along with the two formal and conceptual definitions and representations of an ememe, constitute the major novelty of the present PhD thesis.

The core difference between the two sets of measures lays naturally in the difference between the two ememe representations. The set of fuzzy measures allows a more precise evaluation of the evolution of ememes, while the crisp measures have slightly lower complexity of calculation.

Chapter 6

Evaluation

The approach proposed in the core work of this thesis is aimed at both identifying ememes in Social Media streams and tracking their evolution in time. As previously outlined in Section 4.1.1, this task differs from the tasks of Topic/Event Detection and Tracking, which have been widely explored in the literature. Due to the novelty of the task approached in the core work of this thesis, there is a lack of baseline methods and datasets in the literature for a comparative evaluation purposes. The prior works on information evolution modeling, tracking and analyzing, presented in Chapter 2, have performed preliminary and specific evaluations. In the core work of this thesis, to evaluate the proposed approach we aim to asses two distinct aspects: the quality of the ememes extracted from a stream of Social Media posts, and a quantitative evaluation of the proposed ememe measures. In Section 6.1 we introduce the datasets we created specifically for this purpose. Then we describe the performed data preprocessing steps (Section 6.2) and the phase of parameter setting (Section 6.3). Finally, in Section 6.4 we present the results obtained by applying the proposed method to the generated datasets. In particular, in Section 6.4.1 we present the results achieved for the evaluation of the crisp-defined ememes and the set of crisp measures, while in Section 6.4.2 we describe the result of the evaluation of the fuzzy-defined ememes and the set of fuzzy measures.

6.1 Datasets Created

In this section we describe the datasets that have been created to evaluate the core work of this PhD thesis w.r.t the tasks of ememes representation (crisp and fuzzy), their identification in Social Media, and the measurements of their properties.

We have chosen the micro-blogging platform Twitter as a case study due to its explicit purpose to spread what people think about the world happenings around them, and to share the posts of the people who they like. For this reason Twitter constitutes a good platform for studying ememes from their identification to their spreading dynamics. Moreover, Twitter exposes some useful functionalities. For example, hashtags constitute an important Twitter mechanism that allows to explicitly specify a topic contained in a *tweet*, i.e. a Twitter post of only up to 140 characters. In the analysis of the information generated by Social Media we are immediately faced with the problem of the immense amount of tweets available and of the number of ememes we can identify. In order to limit the stream of Twitter messages to be analyzed, by maintaining a clear reference to a broad and meaningful topic, we have performed a prefiltering phase of the crawled tweets based on a general hashtag (e.g. `#politics`). Some tweets contain more than one hashtag; this represents either a number of sub-topics addressed in the tweet or a number of different concepts of the main topic. Additionally, the identification of ememes related to a predefined topic allows to decrease some difficult problems typical of NLP, as word sense disambiguation, due to the explicit knowledge of the topical content.

We have created a crawler that collects tweets on three separate topics, using the Twitter Search API 1.1. The three collected datasets consist of only tweets in English language. The topics we have selected are the following: Economy, Finance, Politics. Each topic was crawled with the corresponding Twitter hashtag in the time period between March and April 2016. The statistics related to all three datasets are presented in Table 6.1.

Topic	Hashtag	Number of days	Dates	Number of tweets
Economy	`#economy`	57	01/03/2016 - 26/04/2016	180,690
Politics	`#politics`	48	09/03/2016 - 26/04/2016	303,358
Finance	`#finance`	51	06/03/2016 - 26/04/2016	334,660

Table 6.1: Datasets statistics

6.2 Data Preprocessing

Since in this thesis we focus on Social Media, the data preprocessing step is particularly important. First of all, the tweets arriving in a stream related to a broad topic have been preprocessed by eliminating special characters (except for the common Social Media symbols such as "@" for mentioning a user and "#" for hashtag, and "-" to maintain the hyphen connected word), URLs, punctuation and by performing tokenization. We have not performed stemming and Part-of-Speech tagging; in fact, we have aimed at preserving all possible information in the already short Twitter posts. For the same reason we have used a short stopwords list to filter out only the top most common English words.

From each of the three datasets we have eliminated the respective hashtag by which the dataset was collected along with the word of the hashtag since it is a redundant information contained in each tweet. As part of the Graph-of-Words algorithm we have eliminated all non-unique words per tweet. This is performed due to the fact that each tweet in our approach is represented as a fully-connected graph. Since the words in Social Media can have intentional misspellings, we do not perform any prefiltering or corrections of the words. It is important to notice that the proposed algorithm eliminates non-important information due to the fact that the words that characterize this information are absent in the k-core of the graph of the target Social Media stream. All the timestamps of the tweets have been converted to the UTC+01:00 timezone. Retweets have been preprocessed along with the tweets.

6.3 Parameter Setting

In this section we introduce the experimental setup. In particular, we address the issue of setting the k parameter for the k-core graph degeneracy process.

	Economy		Politics		Finance	
TS granularity	**hour**	**day**	**hour**	**day**	**hour**	**day**
Number of TSs in the dataset	1351	57	1114	48	1145	51
Average number of tweets per TS	133.7	6340.0	272.3	6319.9	292.3	6561.9
Average number of edges per TS	3600.5	65485.7	7164.1	127412.4	8021.7	128987.7
Average number of nodes per TS	647.8	6541.5	1246.6	11303.1	1203.4	10107.4
Perc. of nodes with weight $\geq$ 0.1	**32.6%**	**1.6%**	**9.2%**	**0.55%**	**6.9%**	**0.6%**
Average number of nodes with weight $\geq$ 0.1	210.9	103.3	114.6	62.5	82.5	60.2
Perc. of nodes with weight $\geq$ 0.2	**10.9%**	**0.6%**	**2,7%**	**0.16%**	**2.3%**	**0.23%**
Average number of nodes with weight $\geq$ 0.2	70.3	39.2	33.9	17.8	27.3	23.4
Perc. of nodes with weight $\geq$ 0.3	**5.7%**	**0.33%**	**1.3%**	**0.06%**	**1.2%**	**0.14%**
Average number of nodes with weight $\geq$ 0.3	36.7	21.4	16.1	6.9	14.8	14.4
Perc. of nodes with weight $\geq$ 0.4	**3.6%**	**0.22%**	**0.76%**	**0.04%**	**0.83%**	**0.09%**
Average number of nodes with weight $\geq$ 0.4	23.5	14.1	9.5	4	10	9.7
Perc. of nodes with weight $\geq$ 0.5	**2.7%**	**0.16%**	**0.53%**	**0.02%**	**0.61%**	**0.07%**
Average number of nodes with weight $\geq$ 0.5	17.3	10.6	6.6	2.7	7.4	6.6

Table 6.2: The distribution statistics of the degrees of nodes in all studied datasets. All the values are **averaged** over all TSs in the considered dataset.

Economy		Politics		Finance	
TS = hour	**TS=day**	**TS = hour**	**TS=day**	**TS = hour**	**TS=day**
0.5	0.4	0.3	0.2	0.3	0.3

Table 6.3: k parameter setup for each dataset for each TS granularity.

We have studied the behavior and the influence of the k parameter on the extracted ememes through an experimental evaluation of the proposed algorithm on the different datasets and on different time window sizes. The produced analysis allowed us to identify the desired characteristics in the extracted ememes, based on which the setting of the k parameter was performed. In Table 6.2, for each dataset we present the statistics of the distribution of the nodes' degrees averaged over all TSs, for both time granularities of one hour and one day (for generating the TSs). In Table 6.3 the k values for each dataset are reported, for both granularity sizes of TS.

6.4 Experiments and Results

In the following we will present the experiments performed on the created datasets (see Section 6.1), at two time granularity sizes of an hour and of a day, with the k parameter set as described in Section 6.3 for the crisp (Section 6.4.1) and for the fuzzy (Section 6.4.2) approaches to ememes identification, tracking and analyzing.

6.4.1 Results for Crisp Definitions

In this section we present the experimental results on the identification of crisp-defined ememes and their evaluation with the proposed set of crisp measures. In particular, in Section 6.4.1.1 we illustrate a number of examples of identified ememes with the crisp approach, which was detailed in Section 4.2.3.1. While in Section 6.4.1.2 we present the results achieved by applying the proposed set of crisp measures, defined in Section 5.1. Lastly, in Section 6.4.1.3 we provide a discussion on the achieved results and highlight the main findings.

6.4.1.1 Examples of Crisp-Defined Ememes Identified

In Table 6.4 we present some statistics related to the quantitative aspects of the crisp-defined ememes extracted from our datasets at two time granularities, an hour and a day. From this table it becomes apparent that the larger is the size of the time granularity the smaller is the number of core units of information identified. Additionally, these statistics are descriptive of the crawled datatsets: based on the "Avg. num. of crisp-defined ememes per TS", it is reasonable to conclude that the Economy topic, as represented by the crawled dataset for the studied time period, presents a much larger topical diversity, followed by Politics and, lastly, Finance.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Num. of TSs in a dataset	1,351	57	1,114	48	1,145	51
Avg. num. of crisp-defined ememes per TS	2.4	1.2	1.6	1	1.4	1
Max. num. of crisp-defined ememes in TS	11	3	8	1	5	2
Avg. num. of words in a crisp-defined ememe	5.4	7.2	6.7	10.5	6.8	8.5

Table 6.4: Statistics related to the **crisp-defined** ememes identified in each dataset for each time granularity.

In Table 6.5 we present a number of examples of crisp-defined ememes identified with the proposed approach as the core units of information of the studied Twitter substreams of equal time duration. The Finance dataset was considerably populated by specific terminology and jargons that may have limited the ease of interpretation of the obtained results.

Dataset	TS	Crisp-defined ememes
Finance	10am 25.04	**1.** @clever lend joined @hlpartnership network panel #second #charges #commercial #bridging **2.** touched 25 new #startup #money see 08 gmt time usd #fx #forex news #job #jobs #hiring #business **3.** #stocks #forex **4.** @crowdfundernews help make happen education giving back @indiegogo #education #crowdfunding
Politics	11am 14.03	**1.** @joyonlineghana prez mahama errors 59th indece day brochure avoided **2.** @shaliniscribe #aap #india save more money time discontinue non stop paid advertising propaganda self aggrandisement **3.** anyone know trump political agenda think waiting someone post facebook #sanders2016 #notrump #smh @guardian snooper charter labour threatens hold over bill privacy fears news election setback wake up call merkel media politicians reuters **4.** @bored2tears provoking trump defends supporters rejects responsibility violence rallies **5.** romney campaign kasich #feedly **6.** @dvatw #german #media common claim voters influenced #racism #nationialism bullshit sick ignorance
Economy	8pm 27.03	**1.** china industrial profits grow times #iran #china under president **2.** #warning #us #projected #crash #next 2 3 #months #jeffnielson #centralbanking #induced #depression up @albd1971 **3.** #minimumwage 15 down **4.** add expense look statistics expenditures #android **5.** forget millennials time prepare generation z social media #genz new challenge #jobs **6.** report terror wave adds fears fragile turkish

Table 6.5: Examples of **crisp-defined** ememes identified in one same TS.

6.4.1.2 Evaluations of the Crisp-Defined Ememe Measures

In this section we present the statistics and some visual examples related to the proposed set of crisp measures applied to the crisp-defined ememes identified in the crawled datasets, along with the discussions supporting the obtained results.

All the values in the following tables are averaged per each crisp-defined ememe of TS_n and for all TSs in the dataset, unless stated otherwise.

Crisp Relatedness

In Table 6.6 we present the statistics related to the *crisp relatedness* measure. Additionally

to the averaged *crisp relatedness* scores, we present the percentage of tweets with a zero *crisp relatedness* score per a crisp-defined ememe (% of tweets w/ 0 *crisp relatedness*), and the averaged *crisp relatedness* measure excluding these zero scores (Avg. *crisp relatedness* w/out 0s). The percentage of tweets with zero *crisp relatedness* score ranges from the highest in the Economy case for a day granularity - 72.5%, to the lowest of 44.9% in the Finance topic for a day granularity. The phenomenon of large number of tweets unrelated to a specific crisp-defined ememe in a TS_n presents two interesting aspects. Firstly, the high number of tweets with zero *crisp relatedness* scores correlates with the high number of average crisp-defined ememes per TS, as visible from Table 6.4. Secondly, this phenomenon illustrates the overall diversity of information being spread under a broad topic even in such small granularities as an hour or a day. As visible from "Avg. *crisp relatedness* w/out 0s", in the cases of non-zero *crisp relatedness* the identified crisp-defined ememes are well represented in the tweets. Interestingly, we again observe a correlation between the high averaged number of ememes identified in a TS (see Table 6.4) and the high scores for the "Avg. *crisp relatedness* w/out 0s" - the highest *crisp relatedness* scores, 0.41 for an hour granularity and 0.42 for a day granularity, are achieved for the Economy dataset, followed by the scores for the rest of the datasets and granularities in the range of 0.21 and 0.3. In this context, the observed correlation is plausibly due to the fact that the Economy dataset has more topical diversity, w.r.t. the other two datasets under study, leading to strong topical clusters of tweets. In other words, in the Economy dataset for the *crisp relatedness* measure each extracted ememe is strongly related to some small proportion of tweets in the target TS, whilst completely unrelated to its majority.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Avg. *crisp relatedness*	0.08	0.01	0.08	0.09	0.12	0.14
% of tweets w/ 0 *crisp relatedness*	70.3	72.5	62.7	55.3	57.7	44.9
Avg. *crisp relatedness* w/out 0s	0.41	0.42	0.29	0.21	0.3	0.26

Table 6.6: Statistics on the *crisp relatedness* measure. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

Crisp Coverage

The statistics related to the *crisp coverage* measure are presented in Table 6.7. In accordance with the Definition 5.3 the *crisp coverage* measure is calculated based on the *crisp relatedness*

scores, for which the threshold J is set. As observable from the previous subsection on the *crisp relatedness* measure, for each dataset in this study there is large proportion of tweets per TS that have a zero *crisp relatedness* scores. Thus, after a number of experiments, the above mentioned threshold J has been set to equal to 0.1.

From Table 6.7 it becomes apparent that *crisp coverage* measure presents an inverse correlation with the average number of crisp-defined ememes in a TS_n (see Table 6.4). This result is expectable since the higher is the number of ememes found in a TS_n the less there are tweets of TS_n that cover each one of them. The *crisp coverage* scores range from the lowest in the case of the hourly granularity of the Economy dataset - 13.9%, to the highest in the case of the daily granularity of the Politics dataset - 37.7%.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Avg. *crisp coverage* in %	13.9	21	24.6	37.7	33.2	43.3

Table 6.7: Statistics on the *crisp coverage* measure. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

Crisp Prevalence

Table 6.8 present the statistics related to the measure of *crisp prevalence* of a crisp-defined ememe in a TS_n. The "Crisp prevailing ememe score in %" values are averaged over all TSs considering only the most prevailing ememe per each TS.

Similarly to the *crisp coverage*, the scores of *crisp prevalence* measure decrease with the increase of the number of identified ememes in one TS_n (see Table 6.4 for statistics on the crisp-defined ememes identified per TS). The size of the prevalence of an ememe in a substream with regards to other ememes in that same substream differs from the actual value of the importance of an ememe in that substream by 5.9% in the worst case, which is also the case of the highest number of ememes in a TS_n. In the best case the size of the prevalence of an ememe in a substream is identical to the actual value of the importance of that ememe in that substream. Additionally, it is possible to infer that this difference is larger in the cases of the hourly granularity of the datasets, whilst in the cases of the daily granularity this difference tends to zero. The analysis of the results of the *crisp prevalence* measure leads to a conclusion that in Social

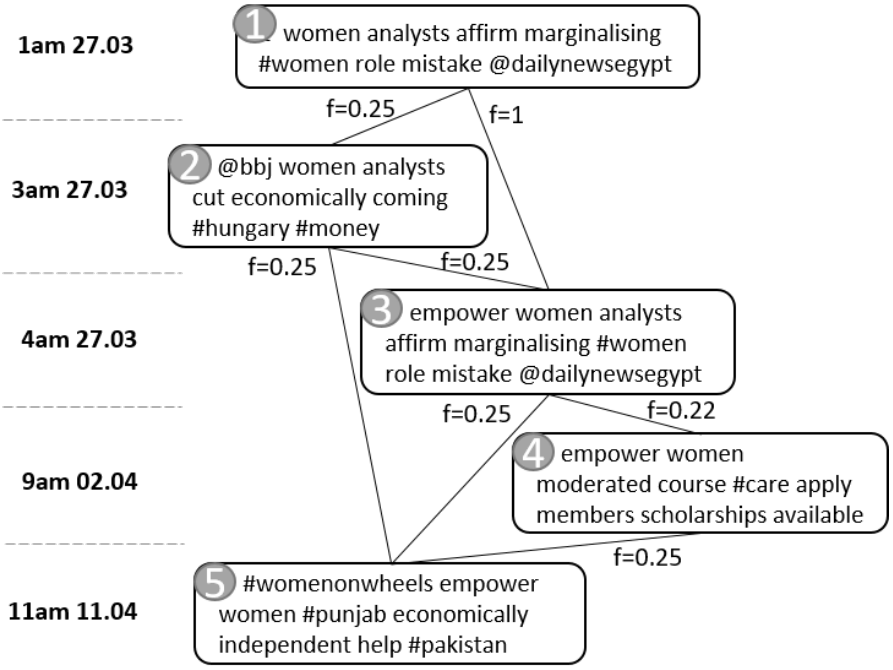

Figure 6.1: *Crisp fidelity* scores between a thread of 5 **crisp-defined** ememes. Economy dataset. TS granularity of an hour.

Media substreams of hourly and daily granularity there is one dominant ememe and a number of weak in importance ememes.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Crisp prevailing ememe score in %	28.9	29.2	40.5	44.5	50.5	54.2
Size of *crisp prevalence* in %	23	28.4	36.8	44.5	45.8	53.8

Table 6.8: Statistics on the *crisp prevalence* measure. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

Crisp Fidelity

In Figure 6.1 we present a visual example of the *crisp fidelity* between a thread of 5 crisp-defined ememes in a period of time of over two weeks. In this figure we present links between ememes in cases they share a *crisp fidelity* score above a threshold of 0.2.

The statistical evaluation of the overall averaged performance of the *crisp fidelity* measure is presented in Table 6.9. The first row of the table presents the averaged *crisp fidelity* (Avg. *crisp fidelity*) between a crisp-defined ememe from TS_n and each crisp-defined ememe from all TSs except for TS_n itself, averaged over all crisp-defined ememes in all TSs for each dataset. Similarly to all the previous measures, the averaged *crisp-fidelity* decreases with the increase of the

number of crisp-defined ememes per TS (see Table 6.4). The next rows present the percentage of crisp-defined ememes with the averaged *crisp fidelity* over 0.1 threshold (% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.1$), 0.5 threshold (% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.5$), and 0.9 threshold (% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.9$), respectively. Accordingly, an inverse correlation between these scores and the numbers of crisp-defined ememes per TS in these datasets is observable. The obtained results for the *crisp fidelity* measure demonstrate an overall coherence between the extracted crisp-defined ememes per each dataset. For example, in the case of the Politics dataset, with granularity of a day, the 100% of the ememes have an averaged *crisp fidelity* over 0.1 to each other ememe from the remaining substreams. This result is closely followed by the Finance dataset with 98% of the ememes above 0.1 *crisp fidelity*, also with granularity set to a day. Moreover, five out of six cases have over 50% of crisp-defined ememes with *crisp fidelity* over 0.1. Setting the *crisp fidelity* threshold to 0.9, the Finance dataset, with a day granularity, achieves the highest result of 29.6%, while the Economy dataset holds the lowest percentage of only 2.7% ememes with *crisp fidelity* over 0.9 to each other ememe from precedent and subsequent substreams. This serves as evidence of the presence of threads of ememes from different substreams with high fidelities in each pair. Additionally, it serves as an example of the difference in behavior of the information spread in different topics - the obviously higher topical diversity of the Economy topic (see Table 6.4) w.r.t. the other two datasets under study leads to a less ememes threads with high fidelity between each ememe in a thread.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Avg. *crisp fidelity*	0.04	0.27	0.28	0.67	0.44	0.71
% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.1$	8.6	55.2	53.3	100	64.3	98
% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.5$	3.9	20.5	27.7	84.4	45.9	81.9
% of crisp-defined ememes with avg. *crisp fidelity* $\geqslant 0.9$	2.7	10.8	8.6	10	21.8	29.6

Table 6.9: Statistics on the *crisp fidelity* measures. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

Crisp Virality

In Table 6.10 we present the various statistics related to the *crisp virality* measure, which is computed as the *crisp fidelity* between the crisp-defined ememe $m_i^{TS_n}$ extracted from the set of

original tweets of the substream TS_n and the crisp-defined ememe $rm_j^{TS_n}$ extracted from the set of retweets of the same substream TS_n. The first row of this table displays the *crisp virality* scores averaged per each ememe $m_i^{TS_n}$. It presents a weak overlap of information between pairs of $m_i^{TS_n}$ and $rm_j^{TS_n}$, especially compared to the *crisp fidelity* scores between ememes from different TSs (for the statistics on the *crisp fidelity* measure see Table 6.9). In particular, the lowest score of 0.06 for averaged *crisp virality* is achieved for the case of hourly granularity of the Economy dataset. Whilst the highest score of 0.37 is achieved for the case of the daily granularity of the Politics dataset. Excluding this last case as an outlier, the average *crisp virality* score over 5 other cases of granularities and topics, we achieve a low score of 0.12. The second row of Table 6.10 presents the percentage of averaged *crisp virality* scores that are equal to zero per $m_i^{TS_n}$ (% of 0 avg. *crisp virality* for crisp-defined ememe). It reaches up to 75.6% in the case of the Economy dataset with an hour granularity, and on average equals to 39% over all cases. Although once again in the case of Politics dataset with a day granularity there is 0% of zero scores for the averaged *crisp virality* measure. The low average *crisp virality* scores and high ratio of zero scores for this measure present an overall incoherence between the ememes extracted from the set of original tweets and the ememes extracted from the set of retweets of the same TS. This suggests a plausible conclusion that the information being spread by a set of original tweets differs considerably from the information being spread by a set of retweets in the same temporal substream. Thus, it indicates that for the replication of information users generally prefer the usage of automated functionalities of Social Media (in this case, retweets) rather than the writing of new original posts. Although, the results achieved for the Politics dataset with day granularity prove the presents of cases in which there is a strong consistency between originally written posts and shared posts.

The last two rows of Table 6.10 present the percentage of average *crisp virality* scores above 0.1 with zero scores (% of *crisp virality* $\geqslant 0.1$) and without zero scores (% of *crisp virality* $\geqslant$ 0.1 w/out 0s), respectively. The *crisp virality* above 0.1 without zeros is on average 64% over all six cases studied, ranging from 43.1% to 90%, while the *crisp virality* above 0.1 with zeros is on average 38.3%, ranging from 11.7% to 83.9%. The large difference between the *crisp virality* with and without the zero scores, provides a strong evidence of a solid coherence between the

ememes from original tweets and the ememes from retweets in those sparse cases when there is a non-zero *crisp fidelity* between them. Along with the previous finding of the concrete case of the Politics dataset with day granularity, these last two rows of the table reveal a presence of very particular cases, regardless of the topic and the size of temporal granularity, in which the spreading of an ememe is performed by both the original tweets and retweets.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Avg. *crisp virality*	0.06	0.13	0.11	0.37	0.1	0.22
% of 0 avg. *crisp virality* for crisp-defined ememe	75.6	31.8	43.6	0	43.8	39.6
Avg. *crisp virality* w/out 0s	0.25	0.19	0.19	0.37	0.19	0.36
% of *crisp virality* $\geqslant 0.1$	11.7	34.4	25.3	83.9	23.5	50.9
% of *crisp virality* $\geqslant 0.1$ w/out 0s	51.7	68.8	46.5	83.9	43.1	90

Table 6.10: Statistics on the *crisp virality* measures. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

Crisp Longevity

The *crisp longevity* measure evaluates the life span of an ememe, which is calculated as the number of TSs in which a crisp-defined ememe is persistent, i.e. has a *crisp fidelity* score over a given threshold. As visible in Table 6.9, all 3 datasets have an overall high inner coherence with the only exception being the Economy dataset with an hour granularity. Therefore, for the measure of *crisp longevity*, we set the threshold referring to the *crisp fidelity* measure to 0.5 for all datasets and all granularities, except for the Economy dataset with hour granularity for which the threshold is set to 0.1.

In Table 6.11 we present the statistics related to the *crisp longevity* measure. The first row presents the *crisp longevity* scores averaged for each crisp-defined ememe per each dataset and per each granularity (Avg. *crisp longevity*). The last row presents the percentage of overall TSs in the dataset that an ememe is *alive* on average (Avg. *crisp longevity* in %). In accordance with the findings of the previous measures, *crisp longevity* is no exception for the inverse correlation with the average number of ememes identified in a TS (for statistics see Table 6.4). The Economy dataset has a higher average number of ememes per a substream and, as visible from Table 6.11, this topical diversity leads to a smaller life span of these ememes, in comparison to the other two topics in this study. Moreover, the results for the *crisp longevity* measure

demonstrate that the crisp-defined ememes extracted at the larger granularity of a day tend to have a longer life span in terms of the number of TSs, with regards to the ones extracted with an hour granularity for all the targeted topics. A possible explanation for this phenomenon relies in the fact that the higher number of posts in a larger TS granularity (day) provides the obtained ememes higher robustness, which translates into a higher resilience over time. Overall, this table demonstrates a relatively large average life span of crisp-defined ememes, ranging for the hour granularity from 8.6% to 45.9% of all substreams and for the day granularity from 20.5% to 84.8%.

	Economy		**Politics**		**Finance**	
Granularity	**h**	**d**	**h**	**d**	**h**	**d**
Avg. *crisp longevity*	285	12.9	494.5	38.8	750.2	41.7
Avg. *crisp longevity* in %	8.6	20.5	27.7	84.4	45.9	81.9

Table 6.11: Statistics on the *crisp longevity* measures. Averaged per each **crisp-defined** ememe of TS_n and for all TSs in the dataset.

6.4.1.3 Discussion on Crisp Results

From the performed analyses it emerged that substreams of Social Media posts tend to have one strongly dominating crisp-defined ememe, and a number of secondary ememes. The proposed measure for evaluating the *crisp fidelity* between two crisp-defined ememes from different substreams indicates an overall high coherence between the extracted temporal sequences of ememes. This offers an evidence of the ability of the proposed method to capture both the evolution of information in time and the presence of ememes that spread relatively intact over time. The proposed measure of *crisp virality* revealed that the information spread by the original Social Media posts is considerably different from the information spread by means of the content sharing functionalities of Social Media platforms, such as "retweets" in Twitter, "share now" button in Facebook, etc. This result demonstrates the tendency of people to prefer the sharing functionalities rather than writing new original post for the specific task of information replication. Although there is evidence of sparse cases in which an ememe is spread with both of the above mentioned approaches for information spreading. Another interesting observation is that the extracted crisp-defined ememes have, on average, a relatively long life span, which

is especially longer for the ememes extracted from the larger time granularity of a day, in comparison to the time granularity of an hour. Overall, the proposed set of crisp measures has not only been able to evaluate the evolution of information as intended, but has also provided valuable insights about different behavior of information and diversity under some topics in comparison to others.

6.4.2 Results for Fuzzy Definitions

In this section we describe the experimental results obtained from the evaluation of the proposed fuzzy-defined approach to ememes identification and evaluation. In particular, in Section 6.4.2.1 we exemplify some samples of extracted ememes, identified according to the fuzzy definition proposed in Section 4.2.3.2. Next, in Section 6.4.2.2 we provide a discussion of the results achieved by applying the set of fuzzy measures to the identified fuzzy-defined ememes, in accordance with the definitions presented in Section 5.2. Lastly, in Section 6.4.2.3 we provide a discussion on the achieved results and highlight some interesting findings of the fuzzy-defined approach.

6.4.2.1 Examples of Fuzzy-Defined Ememes Identified

In this subsection we provide a few examples of fuzzy-defined ememes extracted from the employed datasets, along with some statistics related to them.

In Figure 6.2 we present a sample of the fuzzy-defined ememes extracted from the Economy dataset with an hourly TS granularity; here ememes are represented by means of the words that constitute the support of their fuzzy representations. By the proposed representation it is visible how ememes diffuse and evolve over a timeline of one month. Each of these fuzzy-defined ememes is represented by the most co-occurring words in a tweet over the Economy dataset on hourly basis, as explained in Section 4.2.3.2. In this example, we emphasize with different tonalities in the gray scale and indentation the two initial ememes (the first one in bold black and the second one in bold gray and with indent), and how these evolve over time, gaining

Figure 6.2: Evolution of **fuzzy-defined** ememes in time. Economy dataset. TS granularity of 1 hour.

	Economy		Politics		Finance	
TS granularity	**hour**	**day**	**hour**	**day**	**hour**	**day**
Num. of TSs in the dataset	1,351	57	1,114	48	1,145	51
Avg. number of fuzzy-defined ememes per TS	2.4	1.2	1.6	1	1.4	1
Max. number of fuzzy-defined ememes in one TS	11	3	8	1	5	2
Avg. number of words in a fuzzy-defined ememe	5.4	7.2	6.7	10.5	6.8	8.5

Table 6.12: Statistics related to the **fuzzy-defined** ememes identified in each dataset for each time granularity.

and loosing words in their vocabulary. We leave without any font style and tonality change the words that were only used once and that were not propagated in the next fuzzy-defined ememes.

In Table 6.12 we present different statistics related to the obtained results on the fuzzy-defined ememes extraction from Twitter streams, including the average and maximum numbers of ememes in a TS, and the average number of words in an ememe. In accordance to the crisp-defined ememes results (see Section 6.4.1.1), the statistics for the fuzzy-defined ememes, presented in Table 6.12, demonstrates that the average number of fuzzy-defined ememes per TS is consistently larger (and is greater than 1) for the hourly granularity rather than for the daily throughout all three datasets. This interesting effect of the TS granularity size on the identified fuzzy-defined ememes number is further visible when looking at the vocabulary of the ememes. For example, in Figure 6.2 we observe a 14-ememes long evolution and diffusion thread on the "climate change" and "support local producers" related topic where TS granularity is an hour. Meanwhile, on the same dataset of Economy, choosing the TS granularity set to 1 day, we do not observe a single fuzzy-defined ememe on this topic. Another example concerns the Politics dataset. With the TS granularity set to an hour, 62% of fuzzy-defined ememes mention Donald Trump. While in the same Politics dataset but with the TS set to daily granularity, Trump was mentioned in 100% of tweets. In Table 6.13 we exemplify a few different samples of fuzzy-defined ememes identified with our algorithm in the same TS.

While our datasets of tweets were gathered on three concrete topics, from the obtained results it becomes evident that these broad topics in turn fall into subtopics of numerous ememes throughout the observed datasets. For example, in the hourly partitioning of our Economy

Dataset	TS	Fuzzy-defined ememes
Economy	5am 06.04	**1.** #sales #business prices back **2.** growth global fragile weak imf's lagarde merger news **3.** president **4.** @dollarvigilante watch japan well land rising sun **5.** #android **6.** #money #finance **7.** hailstorm rains lash parts himachal crops destroyed plan #news **8.** rebounding southeast asia seeks buffer china slowdown
Finance	2am 03.04	**1.** @murfo portfolio diversification key maximising return #superannuation #capitalallocation #retirement **2.** 30 great latest #forex #jobs #money #job #hiring #stocks
Politics	2pm 21.03	**1.** trump #tcot obama cuba #news **2.** ireland defends healy rae attacker #news #ireland @whispersnewsltd **3.** political up #news **4.** talk without losing friends #debates #2016election #pointsofview #communication #strategy @strategymgzine

Table 6.13: Examples of **fuzzy-defined** ememes identified in one same TS.

dataset, we find 178 fuzzy-defined ememes concerning "China" spreading throughout the whole observation period. These ememes on "China" concern China's growth, trade, etc. Similarly, we find other countries, such as USA, Greece, Iran, Russia, India, UK, Pakistan, etc. as subtopics, along with common economic matters as production, investing, oil related, stock markets, business, real estate, employment, etc. And all of these "subtopics" concerning countries or economic matters merge and split throughout the whole observation period creating numerous ememes.

It is important to outline that, as observable from Figure 6.2 and Table 6.13, many ememes' timestamps refer to the night and early morning hours. This is due to the fact that, as mentioned in Section 6.2, the timestamps of the tweets were converted to the UTC+01:00 (Europe/Rome) timezone. And since most of the English tweets are geolocated in USA, hence the large time difference and ememes in the middle of the night. This observation is applicable also to the crisp-defined ememes (see Section 6.4.1.1), although not as visible due to the chosen examples.

6.4.2.2 Evaluations of the Fuzzy Defined Ememe Measures

In the following we present the results obtained by applying the proposed set of fuzzy measures on the extracted fuzzy-defined ememes. All the values in the following tables are averaged per each fuzzy-defined ememe of TS_n and for all TSs in the dataset, unless stated otherwise. For comparability of results all the fuzzy-defined measures have been evaluated following the same process as for the crisp-defined measures.

Fuzzy Relatedness

The measure of *fuzzy relatedness* of a fuzzy-defined ememe to a tweet allows us to evaluate how each tweet in the TS_n is related to each ememe, extracted from that TS_n. In Table 6.14 we present the statistics related to the *fuzzy relatedness* measure. In addition to the averaged *fuzzy relatedness* scores we present the averaged percentage of tweets with a zero *fuzzy relatedness* score per each fuzzy-defined ememe of TS_n (% of tweets w/ 0 *fuzzy relatedness*), and the averaged *fuzzy relatedness* measure excluding these zero scores (Avg. *fuzzy relatedness* w/out 0s). The percentage of tweets from TS_n with zero *fuzzy relatedness* score to each observed fuzzy-defined ememe in that TS_n ranges from 44.9% to 72.7%, with the highest numbers for the Economy topic in daily granularity and the lowest numbers for the Finance topic with the daily granularity. This phenomenon of large number of tweets unrelated to a specific fuzzy-defined ememe of a substream TS_n is natural and presents an evident overall diversity of information in the Social Media stream even in a concrete broad topic and such small granularities as an hour or a day. As expected, the *fuzzy relatedness* measure results are from two to five times better when excluding the tweets that did not relate to the target ememe. An observation of Tables 6.12 and 6.14 brings evidence to the presence of a correlation between the percentage of tweets with zero *fuzzy relatedness* scores and the averaged number of fuzzy-defined ememes per TS. This observation is plausibly due to the fact that the more there are ememes identified in a target substream, the more there will be tweets of that substream that are unrelated to some of those ememes.

Overall, the results for the *fuzzy relatedness* measure are in line with the ones achieved for the *crisp relatedness* measure. Although, as visible from comparison of Tables 6.6 and 6.14 it

	Economy		Politics		Finance	
TS granularity	**hour**	**day**	**hour**	**day**	**hour**	**day**
Avg. *fuzzy relatedness*	0.03	0.08	0.04	0.05	0.06	0.09
% of tweets w/ 0 *fuzzy relatedness*	70.9%	72.7%	63.1%	55.4%	58%	44.9%
Avg. *fuzzy relatedness* w/out 0s	0.15	0.29	0.12	0.1	0.15	0.16

Table 6.14: Statistics related to the measure of *fuzzy relatedness* of a **fuzzy-defined** ememe to a tweet.

becomes evident that the averaged scores achieved with the *fuzzy relatedness* are lower than those obtained with the *crisp relatedness*. This result is expectable due to the fact that each word equals to one crisp subset representation of an ememe, while in the fuzzy-defined ememe each word can equal to one only in the best case. Thus, it is logical that the scores for the *fuzzy relatedness* measure are lower, although they are more fine-tuned. Moreover, to take into account the difference between the two ememes representations, the calculation of the two relatedness measures is also different (see Definitions 5.2 and 5.8).

Fuzzy Coverage

Fuzzy coverage measure allows us to identify the overall percentage of the TS_n tweets that include each fuzzy-defined ememes of this TS_n. The threshold J (see Definition 5.9) on the *fuzzy relatedness* score for the *fuzzy coverage* measure is set to 0.1. This value was empirically identified from the statistics illustrated in Table 6.14, where the averaged *fuzzy relatedness*, without the zero scored tweets, is over 0.1 (and equals to 0.1 in case of the daily granularity of the Politics dataset) for all datasets and all TS granularities. In Table 6.15 we present the *fuzzy coverage* value (in percentage) averaged per each fuzzy-defined ememe of TS_n over all the TSs for each dataset. On average, fuzzy-defined ememes were "covered" by 9.5% of TS tweets in the worst case and by 28.9% in the best. As expectable, the Economy topic, that has comparatively the largest number of fuzzy-defined ememes per TS on average and at maximum, as seen in Table 6.12, has the poorest *fuzzy coverage* of fuzzy-defined ememes by the tweets from their TSs, when compared to the other two topics in analysis. This result is in accordance with the findings of the *crisp coverage*, Section 6.4.1.2). As with the relatedness measures, the scores for the *fuzzy coverage* measure are relatively lower than those for the *crisp coverage*.

Fuzzy Prevalence

	Economy		Politics		Finance	
TS granularity	**hour**	**day**	**hour**	**day**	**hour**	**day**
Avg. *fuzzy coverage* in %	9.5%	14.9%	15.2%	18.9%	28.9%	22.6%

Table 6.15: Statistics related to the *fuzzy coverage* measure of a **fuzzy-defined** ememe to a tweet.

	Economy		Politics		Finance	
TS granularity	**hour**	**day**	**hour**	**day**	**hour**	**day**
Fuzzy prevailing ememe score in %	27.4%	28.7%	39.6%	44.5%	49.9%	54.2%
Size of *fuzzy prevalence* in %	21.7%	27.9%	36%	44.5%	45.2%	53.8%

Table 6.16: Statistics related to the *fuzzy prevalence* measure. All values are averaged per each **fuzzy-defined** ememe of TS for all TSs in the dataset.

Fuzzy prevalence measure allows us to understand the importance of each fuzzy-defined ememe identified in the TS_n in comparison to all other fuzzy-defined ememes of that TS_n. In Table 6.16 we present the values (in percentage) of the *fuzzy prevailing* ememe score per TS (*Fuzzy prevailing* ememe score in %), and the gap size (in percentage) between the *fuzzy prevailing* ememe in the TS and the next closest ememe (Size of *fuzzy prevalence* in %).The obtained results demonstrate that the most prevailing fuzzy-defined ememes per TS on average are leading by almost the size of their *fuzzy prevalence*. This result is especially interesting since in four out of six studied cases of the average number of identified ememes per TS was over one, and the maximum number of fuzzy-defined ememes per TS has reached an 11 in the Economy, 8 in the Politics, and 5 in Finance topics, all with hourly granularity (see Table 6.12). This brings to a conclusion, that as a general rule a Social Media substream on a broad topic has one prevailing ememe and a number of ememes with relatively much smaller values and importance levels.

The results for the *fuzzy prevalence* measures are coherent with the ones for the *crisp prevalence* and are with an almost identical magnitude.

Fuzzy Fidelity

Fuzzy fidelity measure allows us to evaluate the persistence in time of the identified fuzzy-defined ememes. In Figure 6.3 we illustrate the *fuzzy fidelity* measure results on a concrete example of 10 fuzzy-defined ememes, from the Economy dataset at daily granularity, concerning a thread of discussion on the raise of the minimum wages to $15 per hour in USA (only values of *fuzzy*

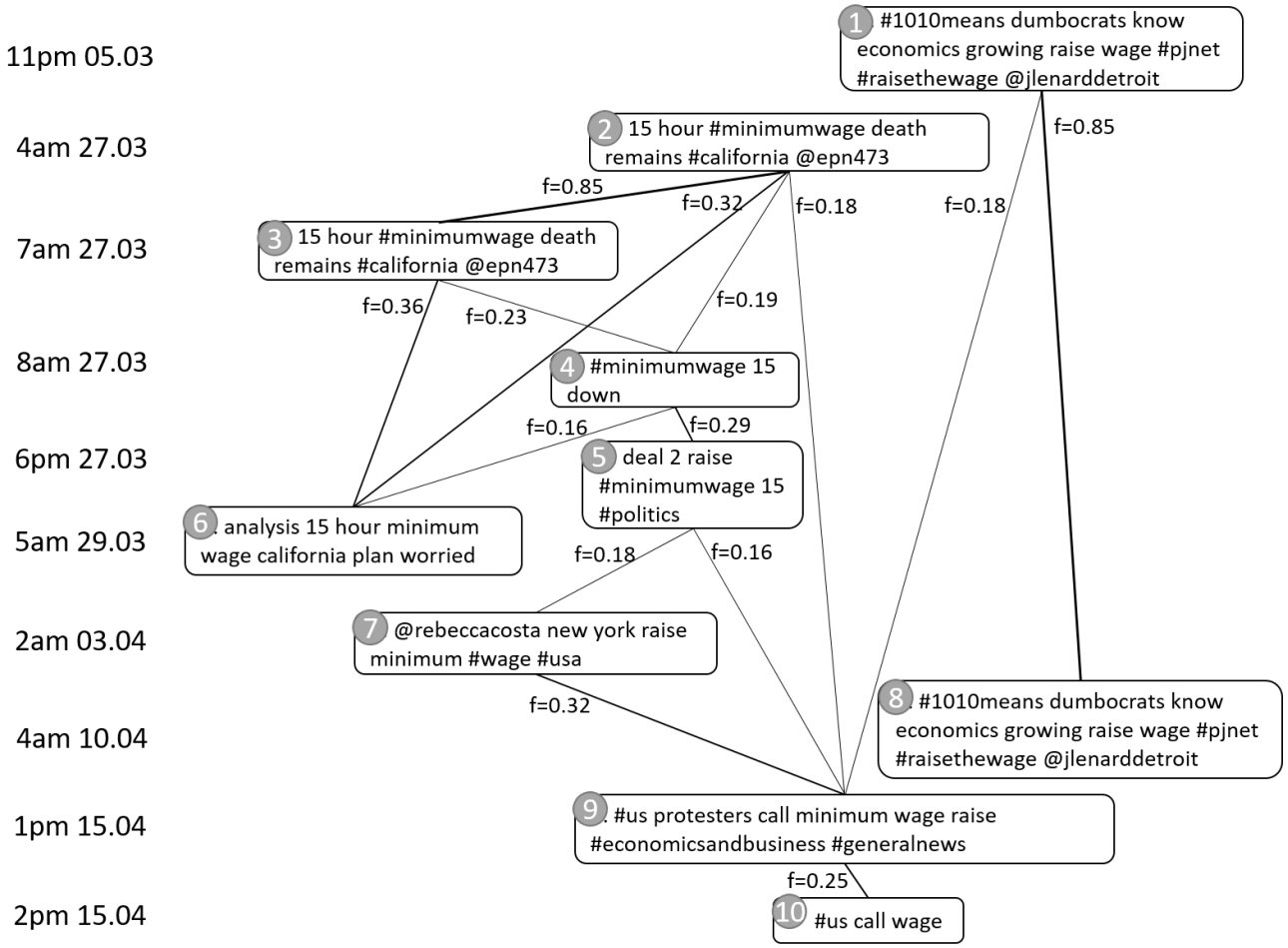

Figure 6.3: Fuzzy fidelity scores between a thread of **fuzzy-defined** ememes. Economy dataset. TS granularity of 1 hour.

fidelity over 0.1 are presented in the Figure 6.3). From this figure we can notice that even though having equal sets of terms, the two pairs of fuzzy-defined ememes 1 and 8, and ememes 2 and 3 have *fuzzy fidelity* of 0.85 due to the fact that the same terms may have different degrees of membership in different fuzzy-defined ememes (see Definition 4.3).

6.4.2.3 Discussion on Fuzzy Results

As observable throughout the whole previous section on the evaluation of the fuzzy approach, the scores achieved for all the fuzzy-defined measures are coherent and relatively lower than those achieved for the crisp measures. As mentioned earlier, this is due to the fact that the membership values of the terms in a fuzzy subset representing a fuzzy-defined ememe may equal to one only in the best case, whilst in the crisp-defined ememe each terms equals to one

due to its presence in the classic subset representing the crisp-defined ememe. Thus, besides introducing a slightly increased complexity for the computation, fuzzy-defined ememes allow a more fine-grained evaluation and understanding of the change in the information through time by permitting to not only evaluate the change of presence in the set of the terms constituting an ememe but also the change of importance of those terms with time. The measure of *fuzzy fidelity* illustrates this issue most vividly, where its scores between two ememes consisting of two identical sets of terms does not equal to the same value due to the fact that thee terms in these two sets have different membership values.

6.5 Discussion

The performed evaluations on three datasets from a popular Social Media platform Twitter demonstrated the potentiality of the proposed crisp and fuzzy approaches to the task of information evolution modeling, tracking and analyzing.

From the point of view of the identification of ememes we believe that both of the proposed methods have presented plausible results in extracting meaningful ememes and in identifying a number of distinct ememes in a substream.

The core difference and advantage of the fuzzy approach to ememes identification and measures is due to a more fine-grained representation of the ememes which allows a more detailed evaluation of their evolution. Therefore, all the measures are calculated differently in the two cases (crisp and fuzzy) to encompass and facilitate the difference in the representations of the ememes, which lead to the difference in the achieved results.

The presented results for crisp-defined and fuzzy-defined measures proved to serve as a good means to track the evolution of ememes in time through the evaluation of the persistence of an ememe in time, its prevalence in a substream, its life span and its representation in a substream by the Social Media posts of that substream.

An important aspect in the tracking and analyzing the evolution of ememes in time relies on

the choice of the granularity for the partitioning of a Social Media stream. Throughout this section we have presented the diversity in the achieved results depending not only on the topic studied but also on the granularity of the time windows. Thus, it is possible to conclude that the proposed crisp and fuzzy measures allow to evaluate the change in time of the information in Social Media but also to understand the behavior of the information streams depending on the topic and the granularity of the analysis.

Chapter 7

Conclusions and Future Work

In this section we present the conclusions and the achievements of this thesis work followed by the future work directions and potential challenges.

7.1 Summary of Thesis Achievements

In this PhD thesis we have addressed the research issue of modeling, analyzing and tracking the evolution of information in Social Media streams in time. Initially, as the first phase, we have explored the potential of Social Media by performing two exploratory studies, which served as the basis for the core work of this thesis. The first exploratory study has put to test the relationship between the polarities of performance of IT companies as expressed in Social Media posts against the Stock Market values of these same companies. The CRF-based classified proposed to use in this study has achieved an average accuracy of 93% for the 8 companies at study. This is particularly interesting since the task of polarity classification is very difficult even for a human being, and especially in such complex topical domains of finance and IT which employ very particular jargons, slangs, symbols and abbreviations. Moreover, the model was built using as features only word combinations and Part-of-Speech tags. Although, the main conclusion of this work is related to the fact of presence of correlation between UGC and the real world happenings, as studied on an example of Stock Market. The second exploratory

work of this first phase of the present PhD thesis concerned the analysis of information spread related to a specific topic, identified under a Social Media hashtag, based on a proposed set of guideline questions and a proposed set of important dimensions. For this second work, we have proposed and developed an Analysis Tool that generates statistical inference about the information spread under a target topic based on different demographic, sentiment and topical dimension of information and their combinations. The thorough experimental evaluation of this tool on a case study has emphasized its potential at extracting both hidden knowledge and unexpected paths that information follows in Social Media. Both of these exploratory studies of the first phase of this thesis have proved the values and importance of the information spread in Social Media in the 21st century. This conclusion laid in the basis of the core work of this PhD thesis, which proposed an innovative methodology for information evolution modeling, tracking and analysis. In particular, the textual information spread on Social Media was targeted as the source of potential knowledge.

The majority of approaches that have been proposed in the literature to address the issue of modeling Information Propagation focus on a graph-based representation of Social Networks in which the information is diffusing [17, 18, 71]. In the core work of this thesis, we have proposed to represent instead the content of a stream of Social Media posts as a graph, by means of a Graph-of-Words representation. Prior works on similar fields have proven the effectiveness of the graph-based representations in the context of Social Media. For the second task of the core work of this thesis - the identification and extraction of ememes from the graph-based representation of Social Media streams - we propose to use an extended k-core graph degeneracy technique. We have proposed two formal and conceptual definitions of an ememe and of its formal representation - the crisp and the fuzzy ones. Most importantly, we have introduced, formally defined and developed two sets of measures - the crisp and the fuzzy ones (for each of the ememe representations), aimed at the qualitative evaluation of the identified crisp-defined and fuzzy-defined ememes and at the quantitative evaluation of the evolution of these ememes in time. Due to the novelty of the tasks approached as the core work of this thesis, leading to an absence of state of art and baseline methods, we have performed an extensive experimental evaluation of the quality of the ememes extracted from a stream of Social Media posts, and

a quantitative evaluation of the proposed ememe measures on three large datasets we have crawled from one of the popular Social Media platforms. This experimental evaluation of our approach has shown its effectiveness in extracting plausible ememes from streams of Social Media posts. The experiments have also outlined the validity of the proposed measures, by bringing interesting insights on different aspects of the information evolution and propagation in Social Media.

7.2 Future Work

There are important and challenging open issues that we would like to address in the future. As one of our future objectives we aim to define a set of additional measures for the identification of events such as birth, merge, split, disappearance of ememes. Another challenging future direction concerns the improvement of the analysis of the texts extracted from Social Media to the aim of modeling the impact of negated words and of the qualification of words through adjectives. Last but not least we also intend to improve the technique aimed at setting the value of the k parameter in the k-core graph degeneracy algorithm.

Moreover, we aim at performing more extensive experiments on diverse Social Media platforms, and on various topics and languages, by also investigating the links of this research to the related fields of Topic/Event Detection and Tracking.

Bibliography

[1] Richard Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, UK, 1976.

[2] Gloria Bordogna and Gabriella Pasi. A fuzzy approach to the conceptual identification of ememes on the blogosphere. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pages 1–8, July 2013.

[3] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[4] Lexing Xie, Apostol Natsev, John R. Kender, Matthew Hill, and John R. Smith. Visual memes in social media: Tracking real-world news in youtube videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 53–62, New York, NY, USA, 2011. ACM.

[5] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM.

[6] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[7] Masaki Mori, Takao Miura, and Isamu Shioya. Topic detection and tracking for news web pages. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 338–342, Washington, DC, USA, 2006. IEEE Computer Society.

[8] Rudy Prabowo, Mike Thelwall, Iina Hellsten, and Andrea Scharnhorst. Evolving debates in online communication: a graph analytical approach. *Internet Research*, 18(5):520–540, 2008.

[9] Hassan Sayyadi and Louiqa Raschid. A graph analytical approach for topic detection. *ACM Trans. Internet Technol.*, 13(2):4:1–4:23, December 2013.

[10] Polykarpos Meladianos, Giannis Nikolentzos, Fran??ois Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Degeneracy-based real-time sub-event detection in twitter stream. In *ICWSM*, 2015.

[11] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event detection and tracking in social streams. In *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009). AAAI*, 2009.

[12] Ekaterina Shabunina and Gabriella Pasi. A graph-based approach to ememes identification and tracking in social media streams. *Knowledge-Based Systems*, October 2017.

[13] Ekaterina Shabunina and Gabriella Pasi. Information evolution modeling and tracking in social media. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 599–606, August 2017.

[14] Ekaterina Shabunina and Gabriella Pasi. Information evolution modeling and tracking: State-of-art, challenges and opportunities. In *Proceedings of the 8th Italian Information Retrieval Workshop, Lugano, Switzerland, June 05-07, 2017.*, pages 102–105, June 2017.

[15] Ekaterina Shabunina, Stefania Marrara, and Gabriella Pasi. An approach to analyse a hashtag-based topic thread in twitter. In *Natural Language Processing and Information*

Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings, pages 350–358, 2016.

[16] Ekaterina Shabunina. Correlation between stock prices and polarity of companies' performance in tweets : a crf-based approach. In *The International Symposium on Web Algorithms (iSWAG)*, June 2015.

[17] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.

[18] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 599–608, Washington, DC, USA, 2010. IEEE Computer Society.

[19] Hector Beck-Fernandez and David F. Nettleton. Identification and extraction of memes represented as semantic networks from free text online forums. In *MDAI 2013 - Modeling Decisions for Artificial Intelligence*, Barcelona, 2013.

[20] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[22] Manuel Gomez-rodriguez, Jure Leskovec, and Bernhard Sch??lkopf. Modeling information propagation with survival theory.

[23] D.J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766, 2002.

[24] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *In SDM*, 2007.

[25] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.

[26] Thomas M. Liggett. *Interacting Particle Systems*. Springer Berlin Heidelberg, 1985.

[27] R. Durrett. *Lecture notes on particle systems and percolation*. Wadsworth & Brooks/Cole statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, 1988.

[28] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, Aug 2001.

[29] Barak Libai Goldenberg, Jacob and Eitan Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, pages 1–18, 2001.

[30] Kyomin Jung, Wooram Heo, and Wei Chen. Irie: Scalable and robust influence maximization in social networks. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 918–923, Washington, DC, USA, 2012. IEEE Computer Society.

[31] Justin Cheng, Lada A. Adamic, Jon M. Kleinberg, and Jure Leskovec. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 671–681, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[32] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Topic-aware social influence propagation models. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 81–90, Washington, DC, USA, 2012. IEEE Computer Society.

[33] Richard Dawkins. *The Blind Watchmaker*. Longman Scientific and Technical, 1986.

[34] Richard Dawkins. Viruses of the mind. In *B. DAHLBOM (ed) Dennett and his Critics*, pages 13–27, 1993.

[35] Daniel Clement Dennett. *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster, New York, 1995.

[36] M. Olesen. *Survival of the Mediated: Speech, the Printing Press and Internet as Selection Mechanisms in Cultural Evolution : Ph.D Thesis*. Faculty of Humanities, University of Copenhagen, 2009.

[37] Ferrante Neri, Carlos Cotta, and Pablo Moscato. *Handbook of Memetic Algorithms*. Springer Publishing Company, Incorporated, 2011.

[38] S Blackmore. *The Meme Machine*. Oxford University Press, Oxford, UK, 1999.

[39] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, New York, NY, USA, 2013. ACM.

[40] Francesco Bonchi, Carlos Castillo, and Dino Ienco. Meme ranking to maximize posts virality in microblogging platforms. *Journal of Intelligent Information Systems*, 40(2):211–239, 2013.

[41] Mohsen JafariAsbagh, Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media streams. *Social Network Analysis and Mining*, 4(1):237, 2014.

[42] Matthew P. Simmons, Lada A. Adamic, and Eytan Adar. Memes online: Extracted, subtracted, injected, and recollected. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[43] Lada Adamic, Thomas Lento, Eytan Adar, and Pauline Ng. Information evolution in social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 473–482, New York, NY, USA, 2016. ACM.

[44] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[45] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[46] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[47] Karen Sparck Jones. Index term weighting. *Information Storage and Retrieval*, 9(11):619 – 633, 1973.

[48] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. pages 275–281, 1998.

[49] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[51] Roi Blanco and Christina Lioma. Graph-based term weighting for information retrieval. *Inf. Retr.*, 15(1):54–92, February 2012.

[52] Tim O'Reilly. What is web 2.0? design patterns and business models for the next generation of software. 2005.

[53] Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. The stock sonar - sentiment analysis of stocks based on a hybrid approach. In *IAAI*, 2011.

[54] Thanh Tien Vu, Shu Chang, Thuy Quang Ha, and Nigel Collier. *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, chapter An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter, pages 23–38. The COLING 2012 Organizing Committee, 2012.

[55] Ilya Zheludev, Robert Smith, and Tomaso Aste. When Can Social Media Lead Financial Markets? *Scientific Reports*, 2014.

[56] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[57] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969.

[58] Yu-Ru Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. #bigbirds never die: Understanding social dynamics of emergent hashtags, 2013.

[59] A. Mislove, S. Lehmann, Y. Ahn, J. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. In *ICWSM*, 2011.

[60] Claudette G Artwick. News sourcing and gender on twitter. *Journalism*, 15 (8):1111–1127, November 2014.

[61] Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 643–652, New York, NY, USA, 2012. ACM.

[62] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: A language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 58–65, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[63] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[64] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of the Advances in Digital Libraries Conference*, ADL '98, pages 12–, Washington, DC, USA, 1998. IEEE Computer Society.

[65] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.

[66] François Rousseau and Michalis Vazirgiannis. Graph-of-word and tw-idf: New approach to ad hoc ir. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 59–68, New York, NY, USA, 2013. ACM.

[67] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, 2001.

[68] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.

[69] V. Batagelj and M. Zaversnik. An o(m) algorithm for cores decomposition of networks. arXiv preprint cs/0310049, 2003.

[70] Wen-June Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85(3):305 – 309, 1997.

[71] Sajid Yousuf Bhat and Muhammad Abulaish. Hoctracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1019–1032, 2014.