UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCES AND ENGINEERING

Analysis of evolution, dynamics and vulnerabilities of Online Social Networks

by

Despoina Antonakaki

PhD Dissertation

Presented

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

Heraklion, February 2020

UNIVERSITY OF CRETE

DEPARTMENT OF COMPUTER SCIENCE

Analysis of evolution, dynamics and vulnerabilities of Online Social Networks

PhD Dissertation Presented

by **Despoina Antonakaki**

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

APPROVED BY:

Author: Despoina Antonakaki

Committee: Research Director, Dr. Sotiris Ioannidis

Committee: Professor, Evangelos Markatos

Committee: Professor, Paraskevi Fragopoulou

Committee: Associate Professor, Konstantinos Magoutis

Committee: Assistant Professor, Polivios Pratikakis

Committee: Assistant Professor, Elias Athanasopoulos

Committee: Assistant Professor, Iason Polakis

Department chairman: Professor Angelos Bilas

Heraklion, February 2020

Acknowledgments

Having completed this journey, I would like to thank everyone who helped and influenced me. First of all, I am grateful to my supervisors Dr. Sotiris Ioannidis and Dr. Paraskevi Fragkopoulou for their initiation of the program that attracted me back to Crete, as well as, for the continuous support and guidance towards my goals. Next, I would like to thank Professor Evangelos Markatos for his guidance and his influence for the realization of my studies and of course this thesis. Also, I would like to thank Professor Iasonas Polakis and Professor Elias Athanasopoulos for their guidance and boost during the beginning of this thesis, as I was entering a new territory and being completely new to cybersecurity. I would like to thank all the remaining members of my committee as well, Professor Polyvios Praktikakis and Professor Kwstantinos Magoutis for the collaboration and their valuable remarks and questions during my defense.

I would also like to thank Panagiotis Ilia, Amira Soliman, Professor Sarunas Girdzijauska and all my ex-fellow students from Marie Curie iSocial ITN program that helped either with comments, or inspirational talks. Particularly, I would like to thank Professor Marián Boguñá and Kaj-Kolja Kleineberg in the "Departament de Física de la Matèria Condensada", of Universitat de Barcelona, during my first secondment, for the discussions and the contribution on the infrastructure. Additionally, I would like to thank Professor Marios Dikaiakos, Professor George Pallis for the collaboration during my second internship in the University of Cuprus, Hariton Efstathiades, Demetris Antoniades, for their valuable comments, as well as the University of Cyprus on the valuable contribution of their infrastructure.

I would also like to thank two older members of the Distributed Computing Systems Lab team, Dimitris Spiliotopoulos and Christos Samaras for the collaboration during the experiments and compilation of the publication "Social media analysis during political turbulence", as well as all the anonymous friends and colleagues who participated in a related crowdsourcing study.

During these years I had the chance to meet and collaborate with a large number of people from the Distributed Computing Systems Lab in FORTH, University of Barcelona, Telefonica and University of Cyprus and whom I would like to thank.

Also, I would like to thank Maria Mamalaki and Euaggelia Kosma, from the University of Crete who helped me during my teaching assistance of the PhD program of University of Crete.

I would like to thank my friends, for their support, love and patience during this period.

There are no adequate words to express my gratitude to my family for the support, help and patience all these years, my brothers Marios, Kostis and Zaxaris and of course my parents Dimitris and Stavroula, who along with my parents-in-law Eleni and Nicolas, took care of my too children, in order to complete this thesis.

Finally, I would like to dedicate this work to my husband Alexandros Kanterakis who is always there for me for 21 years, with support, help, patience and love, as well as, to my two loving children Eleni and Dimitra.

Abstract

Online Social Networks (OSNs) are offering an experience that goes beyond communication, news or entertainment. With a total user base that reaches the one third of the world population and an average daily engagement of three hours, OSNs have become a major phenomenon that affects our society in a variety of ways. Also OSNs have already a history of almost 30 years of constant growth, creating a sizable market that attracts considerable funding and innovation. Inline with this growth, there is a parallel increase of interest from the scientific community that attempts to study OSNs from various perspectives. Without being complete, these perspectives can be delineated according to the way the community treats an OSN as a research object.

First of all, an OSN can be perceived as a complex system represented by a social graph that is continuously changing. A second perspective is as a social phenomenon that hides many dangers from which the public should be informed and protected. A final view of OSNs is as a tool, through which we can focus on some interesting trends and tendencies inherent in the public sphere.

This dissertation presents some fundamental contributions in these areas and uses Twitter as a testbed for experimentation and validation. Initially, we present an effort to model the temporal evolution of the growth of the social graph. Towards this goal, we collect two datasets containing daily snapshots of the social graph, one for the early and another for the later period of Twitter. By fitting this dataset to a well-known but previously untested model, we are able to graph the evolution of Twitter for a period of 8 years. Additionally, we annotate the observed fluctuations of this growth with real events and demonstrate how efficient spam control and service robustness can affect the growth of an OSN.

We proceed to study one of the most common strategies for spam propagation in OSNs. This is the deliberate mix of popular topics with spam content. By using Machine Learning methods, we show that the use of trending topics has the maximum discriminatory efficiency between spam and legit content. Also, we uncover a spam masquerading technique and we show how we can mitigate spam with simple graph analysis and computationally modest machine learning models. Finally, we delve into content analysis. Specifically, we apply a combination of Natural Language Processing techniques to infer how users express themselves during a real and turbulent electoral event. Towards this, we apply Named Entity Recognition, Volume analysis, Sarcasm detection, Sentiment analysis and Topic analysis in order to extract among other, the semantic proximities of different political parties and the temporal sentiment variation of different groups of voters.

Keywords: Twitter, OSN, graph analysis, evolution analysis, spam detection, NLP, Machine Learning, sentiment analysis, topic analysis, classification, sarcasm detection

Περιληψη

Τα σύγχρονα μέσα κοινωνικής δικτύωσης προσφέρουν μια εμπειρία που ξεπερνάει τα όρια της απλής επικοινωνίας, της ενημέρωσης και της ψυχαγωγίας. Με μέσο ημερήσιο χρόνο χρήσης που μπορεί να φτάσει τις 3 ώρες, με μία πληθυσμιακή διείσδυση που ξεπερνάει το ένα τρίτο του παγκόσμιου πληθυσμού, και με ένα σταθερό ρυθμό αύξησης τα τελευταία 30 χρόνια, τα μέσα κοινωνικής δικτύωσης πλέον, επηρεάζουν τον τρόπο με τον οποία μία κοινωνία αλληλεπιδρά, αντιδρά σε διάφορα γεγονότα αλλά και τον τρόπο που διαχέει μία πληροφορία στα μέλη της.

Είναι φυσικό, η τεράστια κοινωνική επίδραση και η επέκταση των μέσων κοινωνικής δικτύωσης, να εγείρει διάφορα ερωτήματα. Μερικά από αυτά, έχουν να κάνουν με τον ρυθμό με τον οποίο μεταβάλλεται και εξελίσσεται ο γράφος που αναπαριστά τους χρήστες ενός κοινωνικού δικτύου και αντιμετωπίζει θέματα όπως, τι αυξάνει περισσότερο με τον χρόνο, οι χρήστες ή οι συνδέσεις που κάνουν μεταξύ τους. Ένα άλλο θέμα είναι η έγκαιρη και αποτελεσματική προστασία των χρηστών από απειλές όπως ανεπιθύμητα μηνύματα. Ένα τρίτο ερώτημα είναι πώς μπορούμε να αποτιμήσουμε την γενικότερη εντύπωση, θετική ή αρνητική, που έχουν οι χρήστες σχετικά με διάφορες ευαίσθητες οντότητες όπως είναι τα πολιτικά κόμματα και οι ιδεολογίες κατά τη διάρκεια μιας προεκλογικής περιόδου.

Η παρούσα διδακτορική διατριβή εστιάζει στο δημοφιλές δίκτυο κοινωνικής δικτύωσης Twitter και επιχειρεί να απαντήσει σε αυτά τα ερωτήματα με την εφαρμογή και εξέλιξη μεθόδων από την περιοχή της ανάλυσης γράφων, τη μηχανική μάθηση και την επεξεργασία φυσικής γλώσσας. Αρχικά, παρουσιάζεται ένα μοντέλο σχετικά με την χρονική εξέλιξη και μοντελοποίηση του κοινωνικού γράφου.

Για το σκοπό αυτό, συλλέγονται δύο αντιπροσωπευτικά δείγματα του Twitter, ένα από την πρώιμη και ένα από την πιο πρόσφατη χρονική περίοδό. Χρησιμοποιώντας ένα γνωστό μοντέλο το οποίο όμως έχει εφαρμοστεί μόνο σε μικρούς γράφους, μελετάμε την εξέλιξη του Twitter, σε μια περίοδο 8 ετών. Επιπλέον, αντιπαραθέτουμε τις παρατηρούμενες διακυμάνσεις αυτής της ανάπτυξης με πραγματικά γεγονότα και καταδεικνύουμε κατά πόσο η εφαρμογή πολιτικών εναντίων ανεπιθύμητων μηνυμάτων αλλά και η εισροή νέων χρηστών μπορεί να επηρεάσει την ανάπτυξη ενός κοινωνικού δικτύου. Στην συνέχεια προχωράμε στη μελέτη μιας νέας στρατηγικής για την διάδοση ανεπιθύμητων μηνυμάτων στα μέσα κοινωνικής δικτύωσης. Ο συγκεκριμένος τρόπος διάδοσης εκμεταλλεύεται τον συνδυασμό δημοφιλών θεμάτων (trending topics) στο Twitter με ανεπιθύμητα μηνύματα.

Χρησιμοποιώντας μεθόδους μηχανικής μάθησης, δείχνουμε ότι η χρήση των δημοφιλών αυτών θεμάτων μας παρέχει τον βέλτιστο τρόπο για τον διαχωρισμό των ανεπιθύμητων μηνυμάτων αλλά και των χρηστών που τα στέλνουν. Επιπλέον, αποκαλύπτουμε μια τεχνική απόκρυψης ανεπιθύμητων μηνυμάτων που διαφεύγει από τους μηχανισμούς ανίχνευσης του Twitter (spam masquerading) και δείχνουμε πώς μπορούμε να μετριάσουμε τα ανεπιθύμητα μηνύματα με απλή ανάλυση του γράφου καθώς και τεχνικών μηχανικής μάθησης.

Η τελευταία πτυχή αυτής της διατριβής μελετάει την ανάλυση του περιεχομένου στο Twitter. Συγκεκριμένα, εφαρμόζουμε ένα συνδυασμό τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) για να μελετήσουμε τον τρόπο έκφρασης των χρηστών και κατ΄ επέκταση των ψηφοφόρων, κατά τη διάρκεια ενός πραγματικού και ταραχώδους εκλογικού γεγονότος. Προκειμένου να γίνει αυτό εφαρμόζουμε τεχνικές εξαγωγής των σημαντικότερων οντοτήτων που περιέχονται στο σύνολο δεδομένων, μελετάμε τον όγκο των μηνυμάτων γύρω από τις οντότητες αυτές και ανιχνεύουμε τα ποσοστά σαρκασμού αλλά και των συναισθημάτων γύρω από αυτές. Με αυτές τις τεχνικές καταλήγουμε στην εξαγωγή σημασιολογικών σχέσεων μεταξύ των σημαντικότερων αυτών οντοτήτων, αλλά και την διακύμανση του συναισθήματος στο χρόνο για τις διάφορες ομάδες ψηφοφόρων.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

> If You're Not Paying For It, You Become The Product.
>
> Richard Serra, Interviews, 1970-1980

Social networks go back to the 1970 with the first attempt by the PLATO system [338] and Talkomatic [342], forming the first generation of social media in 1970s – 1980s. The next generation in 1980s – 1990s introduce the Bulletin board systems (BBS) and Internet Relay Chat (IRC), as operating systems with GUI are emerging. AOL [331] and Windows Live Messenger [344] are born in the generation of 1990s – 2000s and MySpace [336] and Facebook [333] in the next generation of 2000s. Since then the evolution of social media has reached the smart-phone age as well, with Twitter, LinkedIn, Instagram, Snapchat, Viber, Reddit continued by a very long list [343].

Today a large proportion of online activity is happening through Online Social Networks (OSNs). Facebook has reached a user base of 2.3 billion monthly active users [333], Instagram one billion as of May 2019 [334] and Twitter 650 million registered users. The average daily engagement of American users in social networks is more than 3 hours [199], while 45% of Americans, between the age of 18 and 24 years old, are Twitter users [281]. With billions of users, they play a crucial role in information dissemination, raising awareness regarding political and social events, entertainment and personal communication. It becomes crucial to study the usage, as well as the growth and the dynamics of OSNs.

The foundation of trust that these follower-based networks have build with users, has lured the community of malicious users as well; there are serious privacy and security issues. Alongside OSNs have been used for astroturfing, spread of misinformation and fake news. Considering the amount of users that depend on this new media and the importance of the dissemination that is being taking place, the consequences of the misusage can be catastrophic [238]. In April 24, 2013 a fake message in Twitter was enough to cause the market to crash [204]. In April 28, 2017 an ultra luxurious fake music festival was organized, the Fyre festival or "Coachella in the Bahamas", selling tickets at $12,000 apiece, through Instagram [65] leading to a scam compared to a humanitarian disaster. Additionally, Twitter has played a very important role in political and social events like the Arab Spring [340] and in the Occupy Wall Street movement [337]. Also it has been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy and predictions of natural disasters [171].

1.1 Twitter

Twitter is an OSN with a very simple data model. Users post short messages, called *tweets*, that can have no more that 280 characters length (used to be 140 until October 2018) [267]. By default, all tweets are public. Users can opt to receive other users' tweets, an action that establishes a *following* relationship. The collection of tweets that users receive from the users that they follow is called *timeline*.

Twitter is ranked as the 3rd most popular OSNs, by having 650 million registered users [335]. According to Alexa, Twitter is currently ranked as the 8th most popular website of the world [16]. With 330 million monthly active users [32], 100 million daily active users that post 500 million tweets per day [154], Twitter has been established as a very important online media for user interaction and information dissemination.

Twitter stands out from other OSNs from the fact that, although it has a typical OSN structure (users connected to users), it is mainly used for news dissemination [161, 175]. This is due to the fact that, accounts representing public and private institutions, news agencies, public figures, music bands, political parties and other collectives of various nature, flourish in Twitter. These accounts use Twitter as a public bulletin board but also as a medium to create strong ties with the public. The combination of these accounts with those representing individual users, make Twitter a very interesting research object in numerous areas like computer, social, urban and art sciences.

1.1.1 The Social Graph of Twitter

An Online Social Networks can be perceived as an ever-changing graph. In this graph, nodes represent individual users (or accounts) and edges are friendship (or following) relationships between them. This graph evolves through time, as new users are added and form new relationships. A model that describes accurately this variation, can be of extreme importance. It can tell us valuable insights regarding the past of the OSN and can be extrapolated to predict its future. This model can be used even further to help us build optimal sampling techniques [184], recommendation systems [43, 49], measurement of users' influence [56, 224] and understand how information propagates [141].

A study of the growth of the social graph plays an important role in the understanding of the malicious activity, as well. Similar work has been done in detection of sybil attacks [352, 353], spam campaigns [46] and hijacked accounts [302].

1.1.2 Threats on Twitter

Malicious activity on Twitter can be split in two categories based on the incentives of the users or even organizations who initiate them. The first is the distribution of content that directly attempts to get financial gain by tricking unsuspecting users to click malicious links. These links contain either spam or phishing content. The second category has the incentive of misinformation and public deceit and has been publicised as the "fake-news" epidemic. In terms of techniques, both categories use the same basic strategy: reach as many users as possible, as quickly as possible [198]. Maximizing the quantity of victims ensures highest impact, whereas a fast spread minimized the chances of detection and subsequent blocking.

To achieve this dual objective, the offenders, initiate a rapid and mass spread of malicious content, often called "campaigns" [124]. Campaigns are characterized by a large number of accounts that send similar content in an orchestrated manner. Usually this content is interlarded with popular words, keywords or phrases that boost visibility in online search engines. The large number of accounts in a campaign can be either massively created fake accounts or real-user accounts whose credentials have been compromised [302].

1.1.3 Content analysis in Twitter

Online Social Networks are all about content. Whether is amusing, informative, interesting, intimate and empowering, or perhaps infuriating, hateful, distracting, boring, and deceiving is what makes us spent three hours per day in average according to some metrics [15]. Analysis of this content can give us insights into what attracts the attention of people at a given time and place and how do people "feel" about it. The quantification of this "feeling" is possible through a relatively recent technique called Sentiment analysis [134, 244, 246]. "Sentiment" is an attribute that can be assigned to a word, sentence or corpus and can take values "Positive", "Negative" and "Neutral" or more specific values like "Happy" and "Angry". It is based on matching the words of a post with these of a corpus that contains words with predefined sentiment indicators. Additionally, we can generate additional meta-features based on the sentiment values such as "subjectivity" which is the ratio of "positive" and "negative" tweets and "polarity" which is the ratio of "Positive" to "negative" tweets. Sentiment analysis is often combined with Named Entity Recognition which is another Natural Language Processing technique that identifies "real-world" entities (i.e. persons and organizations) in a corpus. The combination of these two methods is commonly referred as "topic analysis".

1.2 Thesis statement and contributions

1.2.1 The research question

This dissertation analyses various aspects of Twitter, contributing thus, in this new branch of science. As with every Social Network, the scientific community is mostly interested in three families of questions:

(1) *How does the social graph change over time? Are there real events that may have influenced this temporal growth?* (2) *What kind of spam do you find in Twitter? Are the spam filtering mechanism of Twitter enough to protect users? How vulnerable are users and how can they get protected?* (3)*What users talk about? How prevalent is a specific topic? What is the general sentiment of users towards a given entity? Is it possible to predict electoral events using OSNs messages?*

These questions are not discrete. Any finding in one question can give a great insight to another. The ecosystem of Twitter is threefold and these perspectives are interconnected. For example a popular discussion topic can be exploited for the distribution of a tweet containing spam. Similarly, the structure of the social graph can be also exploited for the same purpose. Also, the structure of the subset of the social graph that shows positive sentiment towards a given entity (i.e. political party), can give insights of the prospects of this entity. Like living organisms that can be studied from various interconnected perspectives (i.e. physiology, environment, behaviour), Social Networks are dynamic and complex systems that can be studied from many views. This dissertation sheds light in one of the most vibrant social network, Twitter, from three main perspectives: threats, social graphs and content. Initially, we set the ground of this analysis by studying the growth and evolution of the graph of Twitter and then we show how the content is used, as well as being misused.

1.2.2 Thesis

It is possible to use i) follower information of Online Social Networks, such as Twitter, in order to analyze the temporal evolution of the social graph throughout its history, ii) popular trends and messages in order to detect spam, and iii) content information in order to identify user preferences along with semantic proximities of main entities, during political events that attract a significant amount of sarcasm.

1.2.3 Supporting hypothesis

It is possible to obtain sufficient and representative Twitter dataset, parse and collect Twitter messages.

1.2.4 Our approach

- Initially, we conduct a study that implements a fast, efficient and practical method to fit a widely accepted model describing the evolution of the average node degree for large OSNs, in chapter 4. We fit two adequate Twitter datasets to the Leskovec model [184] [186] to measure the temporal growth rate of Twitter and prove that Twitter follows the "densification law". This model states that the average degree of an OSN increases over time. We fit this model on one of the largest samples of Twitter's OSN and we show how it can portray the altering growth periods of Twitter. The dataset consists of two parts: The first part contains all the followers and friends of 92 million users, that we obtained through the Twitter API with the random walk network sampling algorithm [339]. The second part contains all the followers and friends of all users that are present in the study of Kwak et al. [175], which is mainly the entire graph of Twitter, as of July 2009 (40.8 million users).

  In both datasets we use a heuristic [211], to estimate the creation time of user followings, which is not provided by Twitter API. The heuristic is based on the fact that Twitter's API returns the lists of followers and friends of a user ordered according to the link creation time. Next, we sort all edges according to this approximations and we calculate the average outdegree of the network, for every day, between June 2006 and January 2015 (in total 3100 days). The final part is fitting the average outdegree for all days to a pre-existing model ("Leskovec model" [186]), which is a computation task that requires minimum resources. Through this analysis we delineate three types of growth: constant, logarithmic and superlinear. We also calculate a parameter called "growth exponent", which is a single value representation of the momentum of growth for an OSN, at a given time.

- In chapter 5, we perform a comprehensive analysis of the spam mechanism in Twitter. Specifically, spam leveraging trending topics. Trending topics are popular hashtags and popular search queries, which unfortunately consist a very effective method for tricking users into visiting malicious or spam websites, a technique called trendjacking [202]. Initially, we obtain a large dataset of tweets containing popular trends, through Twitter's API. We detect the malicious URLs, contained in these tweets, using Real-Time Blackhole Lists (RBL) and a heuristic to detect obfuscation and label our dataset. In order to separate spam campaigns, we implement a lightweight classifier that relies on specific features of Twitter, while maintaining a very low false positive rate. We show some simple graph techniques to analyze and visualize these type of intense spam campaigns.

- In a period of a bailout referendum in Greece, that took place at 5 July 2015 and the subsequent elections on 20 September 2016, tweet counts could predict the referendum results, without the use of sentiment analysis, considering the demographic subset of Greek Twitter users. Twitter messages in this context were, at least half of them consisting of noisy sarcastic conversations, as mentioned in chapter 6. Additionally, we identify the tweeting patterns, the expressed sentiment and the semantic relations of the most important entities that prevailed during the online discourse that preceded these two events. To accomplish this, we collected all tweets referring to these two events and we applied Named Entity Recognition.

Subsequently, we performed a comparative analysis of the volume and sentiment of the tweets regarding different political parties, politicians and institutions. We also use Latent Dirichlet Allocation, which is an unsupervised learning method that estimates the probability of an entity to belong to a distinct cluster, also called "topic". Each topic is an automatically-extracted semantic structure of the input corpus. This analysis allowed us to extract a 2-D representation of topics, where topics clustered together share the same entities. Hence, we were able to assess the semantic proximities of political parties, politicians and major institutions

1.2.5 Contributions

The main contributions of this dissertation are listed below:

- We confirm that Twitter evolves through time a growing average out-degree although this growth had many fluctuation mainly during Twitter's early period. We make some remarks on several events during the early period of Twitter that may have affected its growth rates. The events are in early phase of Twitter like the first sign-ups without mobiles [329], the SXSW conference [4], the Apple's keynote conference in June 2008, the death of the famous pop artist Michael Jackson on 25th of June 2009 [351], [163] and the blocking from China in early June 2009 [1], have influenced the growth of its graph.

- We prove that the popularity of the messages in Twitter, as indicated in its social graph, increase the misusage of this media. We identify a class of spam that manages to avoid Twitter's spam detection by masquerading spam URLs as Google search results (see figure 5.3), that we call "Gain more followers" campaigns (see chapter 5). We show that Gain More Follower(GMF) spammers are most probably regular users with exploited accounts that interject the owner's legitimate tweets with certain spam tweets, while spammers detected through Realtime Blackhole List [332] have a higher probability of being dedicated spam accounts. Gain More Follower(GMF) spammers use the technique of link farming, as shown in the URLs proved to belong to the Get More Followers campaigns (see figure 5.11).

Our classifier on class spam, achieved a True Positive Rate (TPR) of 73.5%, which is comparable to existing studies, while maintaining a False Positive Rate (FPR) of 0.25% that is significantly lower than existing studies. We also extend the classifier to focus on individual tweets rather than users with similar results (81% TPR and 0.58% FPR).

- We accomplish successful prediction of electoral results during a Greek referendum in 2015. We show that there was decreasing trend of the temporal variation of the ratio of users who included "YES" vs. "NO" entities in their tweets (see figure 6.2). This was in contrast with traditional opinion polls which, according to post-referendum analysis, was erratic [5]. Despite the high difference from the final result (38.6%), the final "YES" vs. "NO" ratio right before the referendum was 18%, which, was very close to the preferences of the demographics of Greek Twitter users. The application of capital controls [330] affected the "YES" vs. "NO" ratio by temporarily strengthening the "NO" sentiment (see figure 6.3).

  The novelty of our approach is that we use a novel sentiment dictionary for the Greek language and that we also account for the presence of sarcasm that has been found to severely confound sentiment analysis [205]. The results of this analysis revealed part of the public sentiment towards main entities along with their semantic proximities. Additionally, we show that there was a strong anti-austerity sentiment accompanied with a critical view on European and Greek political actions.

1.3 Outline of the dissertation

The rest of this dissertation is organized in the following way:

In chapter 2 we summarize the related work associated to the studies conducted in this thesis. Initially, we explain the Twitter features and how Twitter provides content to developers, along with the problems in this area. Additionally, we summarize the background work in the sampling of the social graph and the works related to the generation of time snapshots, which is a process necessary in this thesis. Finally, we report the background work in the analysis of the social graph, the attacks and exploits of Twitter and the sentiment analysis.

Chapter 3 summarized the background work including hashtag and topic recommendation systems and other Twitter features like retweets, mentions, replies and URLs, the social graph of Twitter (PageRank, Homophily) as well as studies about the social graph as a whole, including related works about: "Degree of separations", assortativity and reciprocity. Additionally, related work is included for attacks and exploits including spam, bots and the "fake news" epidemic and finally sentiment analysis, including analysis of languages other than English and psychometric methods.

Chapter 4 presents in extend the study on the average degree and the temporal growth rate of Twitter, analyzing the collection of the data and the generation of time snapshots. Also, this chapter, section 4.4, includes the procedure we followed to fit the "Leskovec model" in our datasets and finally the events at the early stage of Twitter, that we assume have influenced the growth exponent of Twitter.

The next chapter 5 presents the study of the abuse of trending topics and the spam campaigns in Twitter. We analyse the methodology, how the data was collected, the feature extraction , the usage of Real-Time Blackhole Lists (RBL) and the "Get-more-followers (GMF)" campaigns. Finally, this chapter includes the data analysis and the spam classifier.

In chapter 6 we analyze a study on sentiment and topic analysis in a political dataset on Twitter. This chapter includes a subsection on the dataset collection 6.1.1, the entity identification, the sentiment analysis, the new lexicon we compiled for sentiment strength detection and the sarcasm detection. The last subsection of this chapter 6.2 includes the results. Specifically, we present the tweets' volume, the entities co-occurrence, the temporal variation of the sentiment and finally the topic modeling.

Finally, the last chapter 7 includes a synopsis of the contributions and the results of this thesis, as well as directions for future work and research.

1.4 Publications

This section presents the publications in international conferences, workshops and journals that includes part of the work of this dissertation:

- Antonakaki Despoina, Sotiris Ioannidis, and Paraskevi Fragopoulou. "Utilizing the average node degree to assess the temporal growth rate of Twitter." Social Network Analysis and Mining 8.1 (2018): 12. [25]

- Antonakaki, D., Polakis, I., Athanasopoulos, E., Ioannidis, S., & Fragopoulou, P. (2014, November). Think before rt: An experimental study of abusing Twitter trends. In International Conference on Social Informatics (pp. 402-413). Springer, Cham. [26]

- Antonakaki, Despoina, et al. "Exploiting abused trending topics to identify spam campaigns in Twitter." Social Network Analysis and Mining 6.1 (2016): 48. [27]

- Antonakaki, Despoina, et al. "Investigating the complete corpus of referendum and elections tweets." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016. [29]

- Antonakaki, Despoina, Dimitris Spiliotopoulos, Christos V. Samaras, Polyvios Pratikakis, Sotiris Ioannidis, and Paraskevi Fragopoulou. "Social media analysis during political turbulence." PloS one 12, no. 10 (2017): e0186836. [30]

- Antonakaki Despoina, Sotiris Ioannidis, and Paraskevi Fragopoulou. "A survey of Twitter research: Data Model, Graph Structure, Sentiment Analysis and Attacks", Submitted to Social Networks - Journal - Elsevier, Manuscript ID SON-S-17-00006.

Chapter 2
Related Work

This chapter presents the related work in this area starting with a comparison of the most important studies with our thesis. Next, we analyse the Twitter features and present the methods that are required to obtain an adequate dataset from Twitter, in order to conduct a study. Finally, we analyze the related work on the social graph, the attacks and exploits and the sentiment analysis.

2.1 Comparison with related work

The comparison of our analysis with related work, in the studies that we have conducted, is summarized below. The most prominent works, related to our study "The average degree and the temporal growth rate of Twitter" are, as also shown in table 2.1:

- The study of [155] has a smaller dataset comparing to our study: including 87,897 nodes, 829,053 edges from 2007, while our dataset has two parts: 90 million nodes from 2016 and 40.8 million users from 2009, including as well with all friends and followers of these users. Also we provide historical data, while they don't. They measure the user growth, the posts growth rate, which we do not include to our analysis the indegree, and outdegree distributions which we also measure. Additionally, we pinpoint several events in the timeline of Twitter that we believe have affected it growth rate.

- In [230] they provide a dataset of 2012 with 175 million active users and approximately 20 billion edges, but do not report whether they have historical data. They also measure the in-degree distribution, out-degree distribution, and additionally they provide the strongly and weakly connected component sizes, the clustering coefficient, the two-hop neighbourhoods, as well as the shortest path lengths. They do not provide any correlation with events in the timeline of Twitter.

- In [172] they provide a dataset from Yahoo and Flickr, but with no dataset size available and no Twitter data as our study. They build a model of network growth which captures aspects of component structure, they measure the reciprocity, the average and the effective diameter, which we do not provide in Twitter. They do have historical data and or any connection to real events that may have affected this metrics.

- In [186], they provide multiple datasets from networks, not including Twitter: an ArXiv Citation Graph of 29,555 papers and 352,807 edges, a Patents Citation Graph with 3,923,922 patents, an affiliation graphs with 19,309 nodes, an email Network with 3,038,531 emails, an IMDB Actors-to-Movies with 334,084 movies and a product Recommendation Network with 15,646,121 recommendations. This study measures the effective diameter, the average node and the out-degree, the giant connected component and provide a forest fire model graph generator. We actually manage to fit this model into two datasets of Twitter.

The most prominent works related to our study of "Trends and spam campaigns in Twitter" are, as also shown in table 2.2:

- In [139] they provide a complete work on Twitter spam. They obtain a dataset from January to February, 2010 over 200 million tweets from the stream and crawled 25 million URLs. They conduct a complete features analysis, they measure spam click-through statistics and they develop a technique that clusters accounts into campaigns and identify trends that uniquely distinguish phishing, malware, and spam. They examine whether the use of URL blacklists help to significantly stem the spread of Twitter spam. We provide 240 PTs per day and 1.5 million tweets and study the spam mechanism in Twitter, specifically the one leveraging trending topics. Also, we identify a class of spam that manages to avoid Twitter's spam detection by masquerading spam URLs as Google search results. We analyze and visualize these type of intense spam campaigns and build a classifier with (TPR) of 73.5%, (FPR) of 0.25%). We also extend the classifier to focus on individual tweets with 81% TPR and 0.58% FPR.

- In [123] they classify spam campaigns, instead of individual spam messages, they reconstruct campaigns in real-time by adopting incremental clustering and parallelization. They identify six features that distinguish spam campaigns from legitimate message clusters in OSNs. They develop and evaluate an accurate and efficient system that can be easily deployed at the OSN server side to provide online spam filtering. They succeeded 55 % TPR and 0.4 % FPR on Twitter data and 80.8 % TPR and 0.32 % on Facebook data while our average TP rate of the classifier was 81 % (95 % CI [78.5, 82.4]), and the average FP rate was 0.59 % (95 % CI [0.56, 0.62]).

- In [302] they detect hijacked accounts, classify tweets into clusters, classify the clusters, then label each account in the dataset as benign, compromised, or fraudulent.

Table 2.1: The most important works related to our study: The average degree and the temporal growth rate of Twitter

Related Study	Dataset	Findings
Our study: "The average degree and the temporal growth rate of Twitter" [25]	90 million nodes from 2016 and 40.8 million users from 2009, with all friends and followers of these users with historical data	Utilization heuristic to obtain historical data, Fit Leskovec model, The average outdegree, Events at the early stage of Twitter
"Why We Twitter: Understanding Microblogging Usage and Communities" [155]	87,897 nodes, 829,053 edges, no historical data	User growth, Posts growth Rate, Indegree, Outdegree
"Information Network or Social Network? The Structure of the Twitter Follow Graph" [230]	175 million nodes, 20 billion edges, No reference of historical data	In-degree distribution, Out-degree distribution, Strongly and weakly Connected Component sizes, Clustering Coefficient , Two-Hop Neighbourhoods , Shortest Path Lengths.
"Structure and Evolution of Online Social Networks" [172]	Yahoo and Flickr data , no dataset size available	Model of network growth, Singletons, Giant component, Reciprocity, Average and Effective diameter
"In Graph evolution: Densification and shrinking diameters" [186]	ArXiv Citation Graph 29,555 papers e = 352,807 edges. Patents Citation Graph, 3,923,922 patents. Affiliation Graphs 19,309 nodes Email Network 3,038,531 emails IMDB Actors-to-Movies, 334,084 movies Product Recommendation Network 15,646,121 recommendations	Effective diameter, Average node out-degree, Giant connected component, Forest Fire model, Graph generator based on forest fire spreading process

They obtain 8.7 billion tweets originated from 168 million users between January 7, 2013 through October 21, 2013. We are classifying spam campaigns and delve into mechanism of spam masquerading, like GMF campaigns.

- In [301], they study the spamming tools and techniques of 1.1 million suspended Twitter accounts and 80 million tweets posted by them, they study a number of properties of fraudulent accounts (the formation of social relationships, account duration and dormancy periods) and evaluate the wide-spread abuse of URLs, shortening services, free web hosting, and public Twitter clients by spammers. They conduct an in-depth analysis of five of the largest spam campaigns targeting Twitter and finally, recognize an emerging marketplace of social network spam-as-a-service and analyze its underlying infrastructure. We are not targeting specifically suspended account, but initialize our study from unknown set of user and classify them. We detect the spam campaigns that contain a special form of spam: the "GMF campaigns".

The most prominent works related to our study of "Study through NLP and Sentiment analysis in Twitter" are, as also shown in table 2.3:

- Gayo-Avello in [125] is listing number of characteristics and sub-characteristics, defining any method to predict electoral results from Twitter and concluding, as shown in literature, regarding electoral prediction the prediction is not working in Twitter. They list similar work for predictions in elections, as well as, the results indicating poor accuracies. They refer to data cleansing, bias in Twitter data and the weakness of related research. This is probably the most influential work in this area. Our study includes a successful prediction of a referendum, in the limited electoral group of users in Twitter, in Greece, and additionally, we include sentiment analysis and sarcasm detection.
- In [178] they apply sarcasm detection that focuses on user and word selection techniques on a dataset of 60 million tweets, produced by approximately 42K UK Twitter users from 30/04/2010 to 13/02/2012. They train a text regression model that exploits word and user spaces by solving a bilinear optimisation task. We provide two separate datasets of 301,000 and 182,000 tweets for two electoral events, apply sarcasm detection along with sentiment analysis and prediction.
- In [273] they obtain all Dutch tweets of one month from 1,017 randomly selected users, who post messages in Dutch. They collect 1.7 million tweets, out of which they select ones containing names of political parties resulting in 7,000 tweets. They concluded that tweet counting is not a good predictor and can be improved by sentiment analysis. Our study includes a successful prediction, sentiment analysis and sarcasm detection for Greek Twitter data.

Table 2.2: The most prominent works related to our study: "Trends and spam campaigns in Twitter"

Related Study	Dataset	Findings
Our study: "Trends and spam campaigns in Twitter" [27], [26]	240 trending topics, 1.5 million tweets, per day, 150 million tweets totally .	Feature analysis (11 features), Classification of spam tweets and spam users , Average TP rate of the classifier 81 % (95 % CI [78.5, 82.4]), and average FP rate 0.59 % (95 % CI [0.56, 0.62]) Prevalence of trends, Analysis of Get-more-followers (GMF) campaigns.
"@Spam: The Underground on 140 Characters or Less" [139]	200 million tweets	Feature analysis, Spam clickthrough statistics and Clustering spam.
"Toward online spam filtering in social networks" [123]	Facebook: 187 million wall posts, by 3.5 million users January 2008 - June 2009	Classify spam campaigns, Reconstruct campaigns in real-time, Features distinguishing spam campaigns
"Consequences of Connectivity: Characterizing Account Hijacking on Twitter" [302]	8.7 billion tweets 168 million users, Obtained from: January 7, 2013 till October 21, 2013.	Detecting hijacked accounts, Classify tweets into clusters, Classify each cluster as either a benign meme, and labeling accounts as benign, compromised, or fraudulent.
"Suspended Accounts in Retrospect: An Analysis of Twitter Spam" [301]	1.1 million suspended Twitter users , 80 million tweets	Study properties of fraudulent accounts, Evaluate wide-spread abuse of URLs, shortening services, free web hosting, and public Twitter clients by spammers, Recognize emerging marketplace of SN spam-as-a-service

- Skoric et al. in [280], they also report that there is certain correlation between Twitter chatter and votes it is not enough to accomplish accurate predictions. They support that Twitter can provide a good idea on national results but it fails on local levels.

- In [63], they collect tweets from Twitter API reaching 13,899,073 tweets (on 28 November 2014) where they apply sentiment analysis and entity identification. We are conducting these studies, along with prediction of electoral event, sarcasm detection and topic analysis.

- In [83] they study a dataset of 34,697 tweets, collected from January 13 to January 20, 2010 from 2010 US Senate special election in Massachusetts. They report that existing political party classification systems, based on sentiment analysis, are no better than random classifiers. Our study includes a successful prediction, sentiment analysis and sarcasm detection for Greek Twitter data.

- One of the first studies comparing sentiment analysis in Twitter with "traditional" opinion polls was from 2010, demonstrating a strong correlation between Sentiment Analysis in Twitter with Obama's approval ratings polls was [241]. Our study includes a successful prediction of a referendum, in the limited electoral group of users in Twitter, in Greece, as well as sentiment analysis and sarcasm detection.

2.2 Twitter features

Although in principle Twitter is a simple messaging service, users can augment the semantics of a posted text with additional features. These features have contributed immensely to the popularity of Twitter.

2.2.1 Hashtags and trends

Users can enrich the semantics of their messages with terms, called hashtags. A hashtag is a word that is precedented with a hash (#) character, (i.e. "#funny'). These words are indexed separately and users can query the platform in order to find tweets with specific hashtags. Hashtags have evolved to a social phenomenon and their use has been adopted by several media -including non online- as a simple method to signify, idealize and conceptualize a single word (or phrase) in a short message. The general action of assigning hashtags to events, places or people has been described as "social tagging' [153] and is a vital part of Twitter and microblogging in general. Metrics that are applied in hashtags are frequency, specificity, consistency and stability [248]. Popular hashtags and common search terms are listed separately, as popular "trends'. In the literature these are also referred as "topics', "popular trends" or "trending topics". Trends are different per

Table 2.3: The most prominent works related to our study: "Study through NLP and Sentiment analysis in Twitter"

Related Study	Dataset	Findings
Our study: "Study through NLP and Sentiment analysis in Twitter " [29], [30]	301,000 tweets for referendum, and 182,000 tweets for elections	Entity Identification, Compilation of adequate Greek political Dictionary, Prediction of referendum, result Sentiment Analysis, Sarcasm Detection, Tweets' Volume Analysis , Entities Co-occurrence, Temporal Variation of Sentiment , TopicModeling (LDA)
"A meta-analysis of state-of-the-art electoral prediction from Twitter data." [125]	No dataset available	List of features, methods used for prediction of elections in Twitter, Weakness of prediction power of Twitter.
"A user-centric model of voting intention from Social Media" [178]	60 million tweets, 42K UK Twitter users from 30/04/2010 to 13/02/2012.	Text regression model exploiting word and user spaces by solving a bilinear optimisation task.
"In Predicting the 2011 Dutch Senate Election Results with Twitter " [273]	All Dutch tweets: 1.7 million tweets, 1,017 randomly users .	Tweet counting is not a good predictor.
"Can collective sentiment expressed on Twitter predict political elections? " [83]	34,697 tweets	Existing political party classification systems, based on sentiment analysis, no better than random classifiers.

geographic region and the topics that a user views are determined according to the user's friends, interests and location.

The study of Twitter's trends gives valuable information of the importance, duration and impact of real world events. For example, an interesting question is, if Twitter is a fresh content generator, or if it simply reproduces content from external sources. Studies show that Twitter acts as a content aggregator, driving specific trends to popularity [34]. Moreover there is a difference between trends that are emerging from user's activity and traditional headlines that are posted by mainstream media. Analysis has shown that there is a certain qualitative difference between events that are emerging as Twitter trends before appearing in media headlines. Specifically, the events that appear first as Twitter trends, are usually captured by individual users (accidents, demonstrations, happenings, etc.), in contrast to political events that are covered mainly by professional reporters. Finally, 1 out of 5 users tweet about a certain trend and 15% participate in more than 10 topics, within a period of four months [175].

A distinct area of research in Twitter is the semantic analysis of trends and hashtags. The purpose of these studies is to locate semantic relationships between trends and build a trend similarity graph [322]. This graph can help pinpoint emerging topics [231], categorize users into groups of interests [11] and uncover hidden relationships between seemingly unrelated topics [67].

2.2.2 User mentions, URLs, Lists and other features

Another interesting feature is the *User Mentions*. By prefixing a username with the special character @, users can directly refer to a specific user. The referred user is notified about the reference. Finally a tweet can be re-sent, an activity also called *retweeting* and favorited. The number of retweets is commonly associated with the content-value of a specific tweet, whereas the number of mentions is associated with the name-value (or else fame) of the user [70].

Users can also include URLs, pictures or short videos on their posts. Supporting URLs allowed Twitter to play a "briefer" or "bulletin board" role for other online media and content providers who often post a short description of other long posts accompanied with a link to this post. This feature alleviated the text size limitation of Twitter and contributed to the shaping of its conceptual model as a content aggregator rather than a primary content provider. It is estimated that a percentage between 10% [39] and 20% [292] of tweets contain URLs.

A user can also reply to a tweet. In 2010, it was estimated that 23% of tweets got at least one reply [127]. Replies generate a typical thread, commonly seen in online forums.

Another characteristic that is featured are the Twitter lists. A list is created by the owner of a Twitter account, where she can focus on a specific group of people, potentially on a specific topic that should also be indicated by the list name, according to the user's preference. List membership does not involve actions by the involved accounts, but rather by third party users who are curating their own lists and judge two accounts to be similar enough to add to the same list [310].

Table 2.4: Some of the major studies that examine the basic features of Twitter.

	HT	Trends	RT	Mention	Replies	URLs
Metrics	[248]	[175]	[182]	[218]	[127]	[39, 348]
Success	[34]	[31]	[33]	[299, 308]	-	[102]
Influence	[292]	-	[70]	[70]	[350]	[110]
Retweets	[292]	[292]	-	-	-	[292]
Has Graph	[112]	[67]	[39]	[88]	[50, 239]	[39]
Spam	[46]	[203]	[139]	[139]	-	[128]
Bots	[112]	-	[286]	[286]	[112]	[80]

Table 2.4 contains six of the most prominent features of this data model, along with the major research publications that examine them. On the last two categories of spam and bots, underlined cells indicate that the respective studies reached negative conclusions (it is not exploited). Blanks (-) indicate that no study was found that measures the property of this feature. These feature are hashtags (HT), trends, retweets (RT), mention, replies and URLs. *Metrics* refers to the prevalence and the basic statistics of the feature. *Success* refers to the studies that assess whether or not this feature is a "success' indicator of real world entities (i.e. scholar articles, political campaigns etc.). *Influence* refers to the studies that measure the contribution of this feature on the influence of the user. *Retweets* contains the studies that measure if the use of a feature increases the chances of being retweeted. *Has Graph* contains the studies that construct and analyze graphs with this feature as nodes. *Spam* contains the studies that measure the prevalence of this feature on tweets containing spam URLs. *Bots* contains the studies that assess whether this feature has been exploited by bots.

Finally, users can "like' a tweet, although this feature is gradually phased out by the service [227].

2.2.3 Content accessibility

Unregistered users can read tweets, while registered users can also post tweets. Registered users can configure the privacy settings of their profile to render their tweets as public or protected. Public tweets (the default setting) are visible to anyone (registered and unregistered users). A user can configure the privacy settings of her profile to render her tweets as public or protected. Protected tweets may only be visible to users that are preapproved by the original poster. The follower-following relationships in Twitter are not necessarily bidirectional. Any user can follow any other users, in contrast "friendships" in Facebook are established after a mutual agreement between two users. For this reasons a "friendship" graph in Twitter is directed whereas on Facebook is not. The "timeline' of a user contains the temporal updates of the tweets of the users that she follows. Therefore, any user has a set of "followers' (users receiving the tweets that this user sends) and "followings' (users whose tweets appear on this user's timeline).

2.2.4 Getting data from Twitter

Twitter offers an API (Application Programming Interface) for accessing data as well as most of the service's functionality. The access policy to the data was initially very open [216]. The publications regarding Twitter from 2010 [175], indicate that it was possible to collect the entire social graph of Twitter in a period of 2 months, by using only 20 workers. Additionally, Twitter used to "whitelist' IPs with unlimited access for research purposes [46, 119].

In the fear of third-party services misusing the API, in order to build applications that could essentially mimic its main functionality, Twitter started enforcing more strict rate limits [314]. Currently, API requests of authenticated users, or third-party applications, are monitored on a per 15 minute window [313]. In this window, API requests are limited according to their type. For example, in order to request the timeline of a user, a client can perform 900 requests per allowed window. Each request can fetch up to 200 tweets and clients can only fetch up to 3,200 of a user's most recent tweets. Twitter has made available paid data access plans that have more relaxed limits. This is in concordance with Twitter's almost constant effort to monetize its service [304].

To overcome these limitations, researchers, usually utilize multiple applications, risking violating Twitter's Terms of Service and getting their applications suspended. For this purpose, a considerate part of research on Twitter (and other social networks) includes the design and implementation of sophisticated data crawling techniques that respect these limitations, while fetching adequate amount of data [53]. A review of Twitter data collection and survey methods is available from the authors of the Twitter vigilance [69] platform. This platform offers real time influence estimations and sentiment analysis, coupled with a user friendly dashboard. TwAwler [255] can crawl the complete set of tweets, following relationships and other meta-data of an entire community, as big as the Greek (about 330 thousands), by using a single authenticated user and a usual desktop PC. Another crawler that focuses on medium size communities is TwitterEcho [54]. A similar implementation that does not rely on Twitter's API and can access historic data is from Hernandez et al. [147].

Researchers should be aware that releasing Twitter data also violates the Terms of Services [1]. As an effect, large and well-studied Twitter datasets, like the Edinburgh Twitter Corpus [252] and the SNAP dataset [349] are now not publicly available. The unavailabil-

[1] Currently it allows the public release of up to 50,000 tweets per day per user. See term I.F.2.a on Developer Agreement and Policy: https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

ity of public Twitter data has severe effects on the measurement of the reproducibility of current research. Also the absence of "gold standards' has stripped the community with the ability to perform comparison studies. Today, there are two suboptimal approaches to circumvent these limitations. The first is releases of anonymized and heavily processed data [2]. The second is releases that include only the unique IDs of tweets and let researches obtain the rest of the data with their own means. Examples of the latter are, the TREC 2011 Microblog Track [209], the SemEval Twitter datasets [232] and the Stanford Twitter Sentiment Data [133].

Twitter's API returns data in JSON format, with a relatively complex structure. Therefore, researchers show a preference towards NoSQL databases, that natively support JSON structured data, like MongoDB, instead of performing complex conversions required for conventional relational databases.

2.3 Sampling the social graph

Before conducting any study in social networks, researchers have to make a crucial choice: the sampling technique. The size of the social graph of Twitter is in the range of hundreds of millions nodes and hundreds of billion edges. This size makes sampling a necessary prerequisite step, mainly due to (1) API access limitations and (2) extreme computational resources for storage and processing large networks.

Apart from some basic graph measurements (for example average node degree), some of the most informative measurements are in the computational order higher of O(N). This renders these calculations practically impossible on the complete social graph of modern social networks. For these reasons, an efficient sampling technique has to be selected prior to any analysis.

We can make two large distinctions of sampling methods. The first category disregards the user's activity, and focuses solely on the network attributes. The second category takes also into account the user's activity.

Leskovec et.al. [184] applied 10 different sampling techniques to various social networks and measured how well each technique captured the properties of the networks. From all sampling techniques, the two that exhibit better performance were Random Walk and Forest Fire. Random Walk is the sampling method where a node is selected by random and is used as a starting point for a random walk in the graph. Forest Fire is the method where we randomly select a node and then we simulate a fire by randomly burning adjacent edges and nodes. The nodes and edges that are not burned are finally selected. Leskovec et.al. [184] also estimated that a good sampling size should preserve at least 15% of the original size, in order to match the most significant graph properties, such as the average in and out degree.

[2] A list of Twitter datasets and related resources: https://github.com/shaypal5/awesome-twitter-data

Although this was an excellent analysis, it has several practical problems when it comes to modern social networks. The most important is that the authors performed their experiments on large social graphs of their time (2006). At that time, Twitter did not exist and Facebook had approximately 50 million users. A sampling size of 15% is still prohibitive for modern social networks, in terms of computational resources. Another limitation is that they do not take into account other valuable information that might make a node worth of sampling. This is the user's activity and user's influence.

In a later study of 2010 [77], researchers tested sampling methods that combined common sampling techniques, with information regarding user's activity and location. They also used different measurement methods that took into account the ability of the sample to capture "diffusion events'. A diffusion event is the spread of a trend, a URL or a retweet. They concluded that the sampling method that exhibited the lowest distortion from the original graph is the Forest Fire, combined with activity information. They also estimated that an optimal sampling size is approximately 30%. Nevertheless, the validation of this technique on Twitter is an open question. Perhaps the largest sampling experiment that has been performed on Twitter is from Gabielkov et al. [120], which sampled the complete social graph as of 2012. This study concluded that common sampling techniques like Breadth-first search, Random Walk and an alleged unbiased sampling technique, suggested by Wang et al. [320], are all biased towards high degree nodes. Therefore, the optimal sampling technique and sampling size is to a certain extend, an open question.

An orthogonal question for social networks is what is the ideal method to generate artificial graphs, with properties similar to the social graphs. In [187], the authors study 70 sparse real networks and find that a generative model build with 'forest fire' technique burning process, can produce a graph with similar community structure to the real ones.

2.3.1 Generating time snapshots

Twitter does not reveal the creation time of the edges (followings) and therefore it is practically impossible to generate a precise snapshot of the social graph, for a given time. This policy, along with the general restrictions of Twitter's API, has generated criticism, since valuable historic data are extremely difficult to obtain [45]. Nevertheless, Twitter provides the exact time a node was created (user registration) and also the lists of followers and friends of a user, ordered according to the following creation time. These two pieces of information can be combined to produce a lower bound estimation of the following creation time [212]. The accuracy of this heuristic depends on the number of friends or followers of a user. For users with more than 5000 followers, the link creation time is estimated within an accuracy level of several minutes.

Gabielkov et al. [121] performed a temporal analysis of Twitter's macrostructure, by using only the user creation time.

2.4 The Social Graph of Twitter

The social graph of an OSN is defined as the graph on which vertices (or nodes) represent users and edges (or links) represent following relationships. Or else, if user A follows B, this is represented on the graph with a directed edge from node A to node B. Twitter's social graph is directed, which is not always the case in OSNs. For example, in Facebook, a "friendship' is established after a mutual agreement between two users, therefore the formed social graph is undirected.

The social graph has been the center of attention in many research areas. Various properties of this graph are indicative of the nature of the social network and portray how users perceive the platform and interact with other users. It can also provide insight on the temporal dynamic of the platform and the well-being of the service.

Here, we separate studies on the social graph on Twitter in two major categories. The first category studies the social graph at the node level, trying to infer methods that measure the influence, popularity and the social impact of individual users. The second category studies the social graph as a whole, trying to understand the structure and the high-level dynamics of the network.

2.4.1 The Social Graph at the node level

There are two areas of studies that focus on the node level of the Social Graph. The first area studies measurements of the user's influence and the second addresses the phenomenon of *homophily*.

User Influence

In the early stages of Twitter, reaching a high number of followers was considered a strong indication of a user's influence. In one of the most cited papers regarding Twitter, it was shown that the number of followers (*indegree*) is not related to the number of retweets and mentions that a user receives [70]. Indeed, events that require active user engagement (like retweets and mentions) are better estimators of a user's influence, compared to passive followings. Moreover, a tweet from a user with low number of followers, can reach orders of magnitude higher number of users through retweets [182] (the number of users that a tweet finally reaches is also called *impressions*).

These findings sparked an interest for metrics that can give more insightful indications of a user's influence, by also taking into account the user's position in the social network (or else user's topology). Two of the most known metrics of this family are PageRank and Betweenness Centrality.

Betweenness centrality Betweenness centrality is the ratio of all possible shortest paths that pass from a certain node. A node with betweenness centrality equal to 1.0 means that it exists in every shortest path, between any two random nodes of the graph, indicating a maximum influence. This metric, although computationally expensive, can be fairly approximated by sampling a small number of nodes, in artificial networks [38]. Despite this, it is very sensitive to noise since an extra node can alter significantly its value. A variation of betweenness centrality that is computationally lighter and more robust to noise is the K-Betweenness Centrality, which takes into account nodes that lie at most k edges away [194]. A specially designed system for measuring Betweenness centrality on Twitter is GraphCt [106], which employees this metric in order to locate key users for a given topic.

Betweenness centrality belongs to a large collection of measures that try to assess the importance of a node in the network, with information solely from the network topology. Other interesting measures in this family are the closeness centrality (average shortest path length with all other nodes), eigenvector centrality [195] (a predecessor of PageRank) and Katz centrality. Katz centrality resembles PageRank with the fundamental difference that it takes into consideration all nodes of the graph (i.e. not only adjacent nodes), with a weight that is exponentially reduced according to distance [143]. An interesting variation of Katz centrality has been used to assess user influence, by also taking into account the temporal flow of information in the network [177].

Distribution of node degree

Another important property of social networks is the node degree distribution. In directed graphs, like Twitter's OSN, the degree of a node is the sum of the *outdegree* and the *indegree* property. The *outdegree* is the number of edges with direction outward to the node, whereas *indegree* is the number of inward directed edges. The average node degree is a measurement of the density of the graph and characterizes the amount of user interconnections in the network. The average degree has a significant meaning on the modeling of users' behavior, since it has been associated with Dunbar's Number theory, which states that humans can have a finite number of stable social interactions in the range of 100 to 200 [50, 136].

Most importantly, if the log-distribution of the node degree follows a power law [221], then the graph is a scale-free network [220, 305]. Scale-free networks take their name from their general property to have similar structure to parts of themselves (also called self-similarity). The exponent λ of the power law for most real-life, scale-free networks is a value in the range from 2 to 3. Kwak et al. [175] measured the exponent of the power law for this distribution in Twitter to 2.276 in a study of 2009. A study of the same year which examined the topology of 54.3 million users [270], found that both the outgoing and incoming degree follow a power law, with exponents 1.95 and 2.13, respectively. Nevertheless, a study of 2010, with 41.7 million users [175], concluded that Twitter deviates from other social networks and that the outgoing degree distribution is not a power law. The most recent study [230] with the largest sample size (175 million users) estimated that the *indegree* is best fitted by a power law, with $\lambda = 1.35$, whereas the *outdegree* is best fitted by a log-normal distribution, with $\mu = 3.56$ and $\sigma^2 = 2.87$. Given the plethora of contradicting findings, we can conclude that the elucidation of Twitter's degree distribution is an active research question. Nevertheless, all studies agree that Twitter follows a *partial* power law, if we restrict to users with less than ~10^5 followers. This property also contributes to the "small-world' phenomenon described before.

It is important to note that whether the average in and out degree is a power law or not, has an important practical consequence. The mean of a power-law distribution with exponent $\lambda < 2$ diverges, or else it is not strictly defined [236]. If this is the case for Twitter, there is no point in assessing the average node degree, despite being a very simple and intuitive measurement. Nevertheless, according to a study of at 2011 [39], the average *indegree* (followers) of Twitter was measured at 557.1 and the average *outdegree*(friends) was 294.1. For the same year, the average node degree of the undirected social graph of Facebook was measured at 190 [316].

2.4.2 Modeling the social graph

As with any model that describes natural entities, a well formulated model that generates artificial networks, with properties similar to these of real OSNs, is of extreme importance [187].

The main design principle of a mathematical model for the evolution of modern OSNs is to be able to formally describe the behavior of users, in a way that the structure and properties of the network can be predicted over time [172]. Some of the properties that have been observed in large OSNs are the "rich get richer' property [41], the "small world phenomenon' [164] and the decreasing diameter [185]. The "rich get richer' property suggests that new nodes prefer to be connected with nodes with high degree. This is also known as the "preferential attachment' process. The "small world phenomenon' suggests that the average shortest path between two random nodes in the network is proportional to the logarithm of the network's nodes. The "decreasing diameter' suggests that as the network grows, the diameter decreases over time, suggesting that the network "shrinks' or becomes more dense.

Apart from the theoretical interest, these models can have significant impact on the design of practical tools. Examples are sampling techniques [184] and following recommendation systems [43, 49]. A concise model can help build effective defenses against attacks like bots, fake accounts [112] and spam campaigns [46]. Additionally, the area of community detection [42] and measurement of users' influence [56, 224] rely heavily on these models.

One of the latest and most widely accepted model is from Leskovec et.al. in 2007 [186]. This model challenged the existing belief that OSNs evolve with a constant average degree and a slowly growing diameter. In contrast, the authors suggested that, as new nodes are added to the graph, the average degree of modern OSNs is increasing, whereas the diameter is decreasing. This model has been extended [165] to incorporate the layer of the existing, yet unobserved, off-line social network. Although, this model has been validated in 70 small real-life social and information networks [183], efforts to validate it on larger social networks like Facebook [37] and Twitter [25] are partial and inconclusive. The main reasons for this are the limits of Twitter's data access API and the large computational requirements of the validation methods. To put this in a perspective, the computational complexity of measuring the diameter property is in the order of $O(|V||E|)$, which in the case of Twitter, can reach the prohibitive amount of 10^{20} calculations for a sufficient sample size. Also Leskovec et.al. [186] acknowledges that modeling the diameter remains an open question.

Gabielkov et al. in [121] present an effort to create a macrostructure of Twitter. This study, identified the Largest Connected Component (LSS) of the social graph and grouped it as a single node. Subsequently, through breadth first search, they identified smaller components, targeting or being targeted by the LSS. Overall, this technique allowed not only the elucidation of Twitter's macrostructure, but also the exploration of the main patterns of information flow in the graph.

2.5 Attacks and exploits

Generally, social networks have been the target of a variety of malicious attacks [198]. Here, we discuss two of the most serious categories of attacks, with high prevalence on Twitter. The first is spam and the second is automated activity from bot accounts, with the purpose of spreading misinformation and deceit.

2.5.1 Spam

Spam is generally defined as the irrelevant electronic messages in the form of emails, tweets, instant messaging, Usenet newsgroup spam, etc., posted over the Internet. The targeted group includes a large number of users in order to promote, advertise services or products and lure users to malicious activities, malware, phishing etc., although some people define spam generally as any unsolicited mail [341].

Since its early period, Twitter has had a major problem with spam and phishing URLs. A very thorough study of 2010 [139], estimated that, approximately 8% of the URLs posted in Twitter belong to one of these categories. Like in any other social network, spam on

Twitter has two main properties. The first is that, it is usually delivered and spread in the form of massive and orchestrated campaigns [124]. The second is that, spam URLs are most of the time posted from compromised (or else hijacked) accounts [302]. Therefore, two of the main research questions of this area are: Can we predict if a tweet contains spam, or if an account belongs to a spammer? What techniques do spammers employ, in order to maximize the spread of their campaigns?

Workflows for spam classification

Spam detection and classification are the most vivid areas on Twitter studies. There are three types of classification tasks: (1) detecting spam tweets, (2) detecting spam users (spammers) and (3) detecting spam campaigns [81]. If we include bot detection and sentiment classification, we realize that classification is a vital part of many Twitter studies. In table 2.5, we show the general structure and main steps, commonly found in a classification workflow that uses Machine Learning methods. On the remaining of this section, we survey in detail these steps for the task of spam classification.

Labeling data as spam/Legit The first part of a workflow for spam classification is the collection and labeling of tweets, according to their spam/legit status. This dataset will be used to train the classifier and assess the efficiency of the resulted classification algorithm.

Regarding collection, researchers can simply use Twitter's API to collect as many tweets as possible, expecting to "harvest' a fair amount of spam. Interestingly, this procedure can be speed up by building "social honeypots', where multiple legitimate accounts are set for the purpose of attracting and investigating spam and phishing URLs [179, 180, 288].

Regarding spam labeling, one of the most common methods is by employing human inspectors [46, 115, 318]. Although the false positive ratio of this method is very low, an obvious disadvantage is that it requires a considerable amount of human effort. Twitter itself provides the ability to any user to report a tweet or account as spam. This method utilizes the power of the social network itself, nevertheless, reporting data have never been released by any social network.

Perhaps the most efficient method for spam labeling is through the automatic profiling of posted URLs. Due to the restricted size of messages, all URLs in Twitter are shortened to reduce their size, which also has a negative side-effect: the website that the URL points to is hidden and the user only sees the address of the shortening service along with a random identifier. Twitter has employed a URL shortener service that pre-emptively checks for reported malware and phishing sites before shortening a URL [312]. According to a Twitter report from 2010 [78], this service contributed to a drop on spam from 8% to 1%. When users click a URL posted in Twitter, they are either redirected to the initial posted URL (in case of legitimate content) or redirected to a page informing them this URL has been flagged as malicious. This is convenient because researchers do not have to employ any sophisticated technique in order to examine the legitimacy of a link (as for example in traditional mail spam). In contrast, they only have to inspect the response of the URL shortening service when it is asked to un-shorten a URL. Example of studies that use this method for spam detection are [20, 123].

Another method to check the validity of a URL is to query online blacklisting services [20, 203]. Services that have been used in Twitter are PhishTank [173] and Google Safebrowsing [124]. The major drawback of these services is that they exhibit a significant delay (it can be up to 3 days) for updating with novel malicious URLs [276]. Since Twitter is notorious for spreading information rapidly, this can be a major issue. Spammers are aware that domain blacklisting is a very efficient defence mechanism. Specifically, only 2% of spam originates from a dedicated registered domain [301]. For this reason the majority of spam URLs originate from free sub-domains, such as co.cc or dot.tk. These domains cannot be blacklisted as they may contain any kind of content, including legitimate, and they do not have any registration fees. Similarly spammers often exploit free blog hosting services. In the same study [301], the authors revealed that the third most popular domain containing spam is blogspot.com. Other popular choices are LiveJournal and Wordpress. As a conclusion, domain blacklisting should be avoided, as an inefficient strategy, compared to the more targeted approach of URL blacklisting. Examples of domain blacklisting services, that have been proved inefficient, due to delayed updates and containing many false positives, are URIBL and Joewein [139]. Similarly, traditional mail spam defense mechanisms, like Real-time Blackhole List (RBLs), are also ineffective. Another method for evading detection is multiple redirects. A spam URL that goes through multiple redirecting services and lands either in a free subdomain or even better in a blog hosting service, is the most stealthy approach.

Feature Extraction Features for spam classification are account based, like the longevity of the account, the number of posted tweets, the average tweets per day, the number of followers, the number of following and whether or not the account has a description [81]. Some meta-features are the ratio of followers versus followings and the number of bidirectional friends. Features based on tweet content are tweet length, number of URLs posted, number of unique URLs, number of total and unique user mentions, number of trending topics and number of retweets and hashtags. URLs seem to be a valuable source of information for malevolent content. URL features include length, number of subdomain, number of redirections and age of the landing domain [12]. Also, language features include n-grams, similarity of texts sent (a high similarity indicates that the account is actually a robot), similarity of the usernames of a user friends [288] and similarity between the posted Trends and text [20]. An interesting feature is the number of results returned from a web search of an account's name [115], since fraudulent accounts rarely have a web presence. A Facebook specific feature is the user interaction graph [123], which targets users that unexpectedly interact with a high number of friends.

Usually studies extract a subset of the aforementioned features. There is though a distinction between studies that are based on content based features (tweet text or user's profile) and graph based features (based on the properties of the social graph). Papers that belong to the first category are [20,46,180,203,318] and to the second are [20,46,180,318].

It is also possible to measure the classification ability of each feature and rank them accordingly. Available methods for this purpose are the information gain, the chi square [46] and the area under the curve. Another option is to perform classification, by using a single feature and then measure this feature's accuracy.

Classification and Clustering methods After data collection, labeling and feature extraction researchers usually feed this data to a Machine Learning algorithm and attempt to build a SPAM vs. LEGIT classifier. Various algorithms have been tested for this purpose, including Naive Bayes [318], Decision Trees [123, 203], Random Forest [180], Support Vector Machines [46] and Aggregate methods [20]. Some studies also perform unsupervised learning (clustering), with the purpose of generating clusters based on the content [20,123,302]. Towards this direction, a very useful algorithm is the minhash [58]. Clustering helps grouping tweets, and significantly speeds the effort of identifying spam campaigns, in collections of billions of tweets.

The comparison between these studies is not easy, due to the fact that are performed in datasets collected and labelled with different methods. This brings forward the necessity for a publicly available and pre-labelled Twitter spam dataset, similar to various email spam datasets available online [90].

Common spam techniques and practices

Compared to traditional email campaigns, spam in Twitter, appears to follow a more orchestrated and organized approach. Specifically, email spam relies on the bulk distribution of content towards random emails, usually harvested from web crawlers [170]. This is reflected in the clickthrough rate, which is the percentage of spam links that users are tricked to follow, over the sum of the total spam that they receive. For Twitter, this rate has been estimated to be 0.13% [139], which is orders of magnitude higher than the clickthrough rate of mail spam, estimated at 0.01% [160]. In this section we investigate some common exploits used by spammers for rapid content delivery.

Spam content

Another interesting question is *what* is the content that spammers try to promote? First of all, it is very common besides spam, to also post legitimate content in order to avoid detection. Regarding spam, access to entertainment content like music, games and films is ranked as number one [139]. Interestingly content that is most often seen in email spam like pharmaceutical drugs, diet products and adult content is ranked low (less than 5% in total).

Another consideration is the rise of a fraudulent account trading marketplace, that offers additional followers, or even the delivery of thousands of accounts. These campaigns, also called "Gain More Follower' campaigns, attempt to attract victims by offering a mass increase to user's followers. This account selling market generates $127,000 —$459,000 revenue per year [303] just by selling Twitter accounts. The same market also offers accounts for other services like Hotmail, Yahoo and Gmail. Additionally there are rough estimations of the get-more-followers type of spam, that approximate their revenue to multi-millions of dollars [251, 289]. Spammers in these campaigns follow a stealthier approach, compared to other spammers, as they manage to masquerade the malicious URLs behind legitimate and popular sites such as links to Google search results [28]. The revenue of the account selling market is small compared to pharmaceutical drugs promoting campaigns, which is estimated to have a value of 185$ millions [208] or to fake anti-virus markets with a revenue of $130 million [287]. Nevertheless these markets might require to have an actual physical infrastructure (despite selling fake products), compared to the account and get-more-followers markets that requires only the exploitation of account verification mechanisms of the social networks. Other popular spam content is Weight Loss Supplements and Survey Leads [302].

2.6 Sentiment analysis

One of the most promising methods for analysis content in social media is sentiment analysis [129, 201]. "Sentiment' usually is a variable that can take values like: "Positive', "Negative' and "Neutral', or more specific values like "Happy' and "Angry'. Each variable can take a long range of values, allowing for multiple assignments of sentiment in a single word. This means that a word can have both positive and negative sentiment. Moreover, we can generate additional meta-features based on the sentiment values. These are "subjectivity' and "polarity'. Subjectivity is the ratio of "positive' and "negative' tweets to "neutral' tweets. Polarity is the ratio of "Positive' to "negative' tweets. Sentiment analysis can portray the attitude of the public towards a specific issue or the "positiveness', as a personality trait of a single user.

2.6.1 A common sentiment analysis pipeline

The usual methodology for sentiment analysis [166, 244] requires the pre-processing and extraction of lexical features from tweets. Preprocessing includes tokenization and the removal of stop words and other elements, without lexical value like URLs and mentions. Useful lexical features include word stems, Part of Speech (POS) tags [95] and n-grams. Fortunately, there are mature and efficient tools that perform these tasks, with minimal programming effort. Examples are NLTK [47] for Python and MALLET [207] for Java.

Among the lexical features are emoticons and emojis. Studies show that tweets with positive emoticons are 4 times more likely than tweets with negative [284], so researchers need to correct for this imbalance. Regarding emojis, the Unicode standard contains 2,823 emojis and more than half of Instagram posts contain at least one [98]. Research has shown that most used emojis convey both positive and negative sentiment [75] and are valuable features for sentiment detection.

Other popular features are topics and entities. Topics represent clusters of common words, that appear in a set of documents [150]. A document can belong in multiple topics and topics do not have to have a "real world" interpretation. The most common method for topic modeling is Latent Dirichlet Allocation (LDA) and a commonly used implementation for Twitter data is Twitter-LDA [356]. In contrast to topics, entities are notions with "real world" meaning. The task of Named Entity Recognition (NER) is the extraction of a generic semantic identity for a word. For example "Person" for "Obama" and "Place" for "New York". A popular NER tool is Stanford NER [114], whereas T-NER [263] and TwiNER [188] are optimized tools for Twitter.

The manual labeling of sentiment in tweets is done through two possible methods. This first is through a panel of experts and the second is with crowdsourcing techniques. The crowdsourcing technique is the use of online platforms that allow anyone to manually label the tweets, usually with a small reward. Popular choices are CrowdFlower and the Amazon Mechanical Turk [113]. A very early (2008) analysis [283] argues that, expert employment and crowdsourcing techniques produce both, equally qualitative results. A later study (2016) [226] discovered that the quality of manual labeling is more important than the choice of the classification method. A metric that is commonly used to measure the concordance of labeling among multiple workers is the Fleiss' kappa [260]. Interestingly, one method for locating and encouraging users to participate in a crowdsourced dataset labeling task is through Twitter bots [18].

The result of all this pipeline is the construction of a feature rich dataset, that contains linguistic features and sentiments for the collected text from social media. This dataset usually is structured as a $T \times F$ vector space with T being the number of texts and F being the number of features. Alternatively, the extracted features can be modelled as graphs, by importing information from the social graph, via a method called "label propagation" [284, 293].

This dataset can be used in a variety of methods. The first is to show the temporal variation of sentiment, over a course of a specific event. For example, we can visualize the variation of the sentiment of the public, over the course of a certain political campaign.

We can also quantify how specific actions or events altered the public sentiment. Another line of work is to build a machine learning classifier that predicts the sentiment of the public, based on the linguistic features. This can help to quickly assess the sentiment, based on linguistic features and find which linguistic features are more associated to sentiment. One of the most commonly used tool that provides most of the presented functionality is Vader [131], which according to its authors outperforms even human annotators. In table 2.5 we present a typical classification workflow in Twitter, for a variety of classification tasks including sentiment.

2.6.2 Milestone studies, findings and notes

The first work on sentiment analysis in Twitter was performed by Go et al. in 2009 [133]. This work used emoticons as sentiment indicators for labeling, used a train set of 1.6 million tweets and a test set of 300 manually labelled tweets. They extracted text features such as n-grams, bigrams, and Part Of Speech tags, and achieved a classification accuracy in the range of 82%. Many subsequent works used this study as a baseline, based on the fact that they also released the train dataset. Instead of emoticons, Kouloumpis et al. in 2011 [168], used hashtags for tweet labeling. Hashtags were manually labelled as positive (i.e. #success), negative (i.e. #fail) and neutral (i.e. #news). Liu et al. [190], noticed that current models use either emoticons, or manually labelled tweets as sentiment labels for classification, and suggested a hybrid system that imports information from both sources.

After these initial studies, we notice two parallel efforts in sentiment analysis. The first is to incorporate knowledge from external resources and the second is to measure the public opinion towards specific entities like persons, events and products.

Bollen et al. at 2011 [52] was the first study to employ an external lexicon, in order to label the sentiment features of tweets and associate their fluctuations with real events of 2008. This lexicon was the extended version of POMS (Profile of Mood States [250]), which contains 793 terms associated with 6 mood dimensions (Tension, Depression, Anger, Vigour, Fatigue, and Confusion). Two studies from Saif et al. at 2012 [271, 272] considered the use of entity extraction services [264] like AlchemyAPI, OpenCalais and Zemanta and added entities in the feature set of tweets. Finally, Zhang et al. [354] at the same year, used an opinion lexicon [99] tailored for product review analysis, to annotate the lexical features used for each entity of interest (they tested on Obama, Harry Potter, Tangled, iPad and Packers).

Examples of open existing dictionaries for NLP purposes and sentiment analysis are (for more see [254]):

1. SentiWordNet [36], a sentiment lexicon of 100,000 English words.

Table 2.5: The major steps of a general classification workflow for Twitter data with Machine Learning methods.

Step 1. Data Collection	
Methods	Twitter API [53, 147, 255]
	Open Datasets (Limited) [245]
	Social Honeypots [179, 180, 288]
Step 2. Labeling	
Labels	Spam/Legit, Positive/Negative, Bot/Real, Rumor/True
Methods	Manual: Human Experts [226]
	Crowdsourcing [113]: Amazon Mechanical Turk, Crowdflower
	Automatic: Spam: labeling from shortening services (t.co) [20, 123] Sentiment: Emoticons [133, 233]
Step 3. Split data in 3 parts: Train, Test and Validation	
Step 4. Feature Extraction [12, 81]	
Lexical	Tokens, Stems, POS, n-grams, stop words, emoticons
Content	URLs, Mentions, Hashtags, Topics, Date
Influence	Retweet, Reply, Like
Profile	Friends, Followers, Date of creation, Description, #Tweets, Date since last tweet
Step 5. Add knowledge from external resources	
Sentiment Analysis	Sentiment lexicons and vocabularies [254]
Spam Classification	Online Black-list services [20, 139, 203]
Bot Detection	Search engine results for the account name [115]
Step 6. Machine Learning	
Methods	Decision Trees, Naive Bayes, SVM, Random Forests, Deep Neural Networks
Implementation	Weka (Java) [142], scikit-learn (python) [249]
Step 7. Estimate accuracy	
Basic Metrics	True Positive, True Negative, False Positive, False Negative
Accuracy Metrics	Precision, Recall, F1, Accuracy [46]
Sensitivity	Area Under the Curve (AUC), Confidence Intervals [234]
Efficiency	Time and resources needed for the complete workflow [139]

2. OpinionFinder, a subjectivity lexicon[3] containing 2.304 words, annotated as positive and 4.153 as negative.

3. Dictionary of English Stop words[4].

4. The Affective Norms for English Words -ANEW- dataset. It contains emotional ratings for 1034 English words.

5. A Google-based Profile of Mood States (GPOMS) [51]. It assigns 6 emotion values (calm, Alert, Sure, Vital, Kind and Happy) in any text, based on the Google's n-gram collection.

6. The CMU Pronouncing Dictionary[5], with pronunciation information for 134.000 English words.

Regarding sentiment measurement towards specific entities, Diakopoulos et al. [96] used 1,820 manually labelled tweets to measure the temporal variation of sentiment, during the broadcast of U.S. presidential debate in 2008. In [33], the authors exploited the sentiment information, in order to predict the revenue of movie after their release. Jiang et al. at 2011 [156] build a model based on 2,400 manually labelled tweets and lexical features to measure the sentiment towards 5 popular queries (Obama, Google, iPad, Lakers and Lady Gaga). This approach has the benefit of accommodating different uses of words, including slang for different entities. Wang et al. at 2012 [319] was the first to build a real-time sentiment monitor, which was tested on the 2012 US elections. It used a model trained on 1,820 manually labeled tweets, and it was also the first study to also correct for humorous and sarcastic content. Finally, Mitchell et al. at 2013 [222] performed the first study that explores differences of expressed sentiment in various geographic regions.

An interesting trend that appeared on approximately 2014, was the use of Deep Neural Networks for sentiment classification [274, 294, 295]. In one of the first studies [295] the authors used 10 million tweets, in order to build word embeddings and achieved an impressing 86% accuracy on the task of positive vs. negative tweet classification.

2.6.3 Analysis of political discourse in Twitter

Many studies have performed elaborate analyses, in order to investigate the behavior of online users, during pre-election periods. The purpose of most of these studies is to generate patterns that distinguish users' or posts' favoritism towards one political party or certain ideology. Here, the main predicament is to generate election predictions, that are

[3] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[4] http://www.ranks.nl/stopwords
[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

close or even outperform public opinion polls [178], to measure approval ratings [241], or to assess public opinion during political debates [97].

A good indicator for a party's success is the tweet volume, given that the correct time window is defined [109]although studies indicate that this is inefficient without sentiment analysis [277]. Concerning sentiment analysis techniques, researchers make use of specially tailored dictionaries with positive, negative or neutral colored words, measuring the occurrence of these words in a rich variety of language properties of the posted text [247], [244] or hashtags [265].

Gayo-Avello [125] lists the major difficulties of this area that need to be addressed, before making Twitter a reliable election prediction mechanism. In brief, these difficulties are noise and demographics. Regarding noise, a huge proportion of election-related Twitter posts are humorous, ironic or sarcastic and do not portray any party (or ideology) inclination. It is estimated that approximately half of collected tweets belong to this category [22, 72]. Filtering out these posts or users is a challenging task and relies heavily on qualitative human-crafted datasets of sentiment vocabularies and pre-classified, "ground truth" samples [152]. Low-quality, human-curated datasets can result in a very inefficient, classification algorithm, as it happened in a sarcasm detection system [137]. Existing studies on sarcasm detection are focusing on user and word selection techniques [178], or are explicitly addressing reliability level of posts, by classifying them as rumors or trolls [193].

Regarding demographics, Twitter users belong to a specific social group, that is not necessarily representative of the whole electorate. Specifically, studies have indicated that Twitter users belong to a certain age [126], social [256] and ideology demographic group and therefore, express a partial opinion of the society at best. A study of 2011 concluded that due to its demographics, Twitter is by far inferior, compared to opinion polls, for elections prediction in the U.S. [126]. Another study reported that existing political party classification systems, based on sentiment analysis, are no better than random classifiers [83]. This indicates that, sentiment analysis methods are in their infancy and that they should be coupled with more sophisticated methods that incorporate rich lexical properties and context indicators, specific to each campaign [277]. Fortunately, existing techniques can effectively assess and correct these biases [256].

The first term of Barack Obama's presidency (2009-2012) coincided with the immense increase of Twitter's user base and its establishment as a channel for personal political expression. As a consequence, one of the first studies that compared sentiment analysis in Twitter with "traditional" opinion polls was from 2010, demonstrating a strong correlation between Sentiment Analysis in Twitter with Obama's approval ratings polls [241]. The application of the same method in 2012 U.S. presidential elections outperformed the public opinion polls [104]. Since then, numerous studies have performed similar analysis in other countries like Austria [178], UK [178] and Italy [104], with varying election procedures and diverse cultural and language dynamics.

The main research strategies according to Gayo-Avello [125] in the area are: (i) classification according to tweet volume and (ii) classification according to sentiment analysis according. Modern studies usually implement combination of these two main strategies [277].

In other studies we see knowledge extraction from the social graph by studying the retweet or mention graph [87].

Other approaches extract knowledge from the social graph, by studying the retweet or the mention graph [88], or by averaging on the predefined ideology of the political leaders that the users follow [135]. The tweet volume is a good indicator for a party's success, given that the correct time window is defined [109], but studies indicate that this is inefficient without sentiment analysis [277].

Sentiment analysis can measure the polarity of Twitter's users, on regard to a specific event or movement [85, 88]. For example, group polarization have been studied on the context of "Arab spring" [326], the Venezuelan president Hugo Chavez [225], the climate change [84, 237], the European refugee crisis [268] and American politics [321].

Chapter 3
Background work

In order to complete the analysis of the related work this chapter extends the previous chapter by presenting the background work on the Twitter features, the social graph, the attacks and exploits and sentiment analysis that is not so close related to our thesis.

3.1 Twitter features

3.1.1 Hashtag and Topic Recommendation Systems

In topic analysis studies, hashtags, are good predictors of the thematic subject of a tweet [31]. This in turn, makes hashtags valuable in recommendation systems, trying to assist users to assign appropriate hashtags to tweets [108]. Similarly, topic recommendation systems attempt to extract topics that are subjective to users' interests and friendships, while being timely and accurate [68].

3.1.2 Retweets, mentions, replies and URLs

Although retweeting is one of the most known features of Twitter, at 2010 it was estimated that only 6% of tweets got at least one retweet [127]. Therefore, a common line of research is to locate the features that drive a tweet to be more retweeted [55, 292]. One of these studies showed that tweets getting more re-tweeted, have similar textual and thematic content [149]. Specifically, tweets with general thematic content (i.e. Christmas), or bad news are more likely to be re-tweeted [234]. Also, Suh et al. [292] showed that URLs, hashtags, number of followers and followings affect positively the number of retweets, whereas the number of past tweets does not have any effect. In another study [174], it was shown that the position of a user in the social graph (assessed by the PageRank metric) is also a crucial factor.

In cases where a tweet contains a URL promoting a future event, the ratio of the retweets before and after the event is a good predictor of its "success". For example, Asur et al. [33] used these metrics to predict the success of movies right after their release. In [110] the author derived accurate predictions of the citations of a scientific paper, based on the number of tweets containing URLs to the online versions of this paper.

The total number of mentions that users receive is associated with the influence of their profile, as a whole, and not with the impact of their individual tweets. For this reason, user mentions are commonly used to measure the "success" of a user, as opposed to events. Examples that are employing the number of mentions, in order to measure the impact of a user profile, is on scholarly publications [278, 299] and election campaigns [151, 308]. The number of mentions is also useful for assessing the success of a paid advertising campaign in Twitter. This is because mentions require active engagement, in contrast to simple views. Twitter estimates that approximately 80% of its users have mentioned a brand, at least once [218].

The reply-network is a graph where nodes are users and edges represent reply events, in a defined period of time. This network is believed to represent user similarities in a higher level than that the typical users-followers network [50]. Another type of reply-network is the reply-cascade tree, which simply represents the discussion thread initiated by a single tweet [239]. The shape of this tree is highly depended on the *indegree* (number of followers) of the root node [239]. The number of replies that a user receives, is also a metric of influence [350].

Given the hundreds of millions of tweets posted daily, it is evident that an active area of research is the evaluation of Twitter as a general purpose web search engine [296]. Indeed, in a study of Dong et al. [102] it was found that URLs posted in Twitter were more relevant and more recent, compared to results returned from common search engines. Also URLs have a big variety of life span, depending on the category of the poster. According to Wu et al. [348], URLs posted by media organizations are short-lived, whereas URLs from bloggers have a longer life span, especially if they link to music or video content. The same study also concluded that 50% of URLs posted in Twitter are "generated" (or else initially posted) from a very low number of "elite" users. In section 2.5.1 we survey another crucial area of research, which is the presence of spam and other malicious URLs in Twitter.

3.2 Social Graph of Twitter

PageRank Regarding topology, the most widely known algorithm that measures a user's influence in a social network, is PageRank [57]. PageRank was initially designed by Google to measure the relevance of a website on the Internet, regarding a search term. To assess the value of PageRank for all nodes, we initially assign small constant (or random) values to all nodes of the graph. Then, for each node, we re-assign this value to the weighted sum of the PageRank of all nodes that link to this node. We iterate this step until the assigned values converge. Once it converges, a node with high PageRank value means, in general, that it has other nodes with high PageRank value linking to it. The application of the same algorithm in social graphs indicates the influence of a user. Studies that have applied PageRank on Twitter include [175] and [327]. Variations of PageRank exist that are more tailored to the context of Twitter, like the Influence-Passivity Algorithm [266]. Also the Hirsh index [148] resembles PageRank, since it is an influence measurement algorithm that also originates from a different concept (measurement of scientific impact) and can be adapted to social networks.

It is interesting that these studies have done limited work in comparing PageRank with other methods, and they limit their analysis on presenting the top popular users on Twitter according to this metric. This might be part of the wider issue, that there isn't any gold standard or a widely accepted methodology for assessing the accuracy of user influence methods. PageRank (and its variations) is utilizing solely the network structure, in order to assess a user's popularity. Nevertheless, we have seen that topic modeling is also important for measuring the influence of a user, in respect to a certain topic. Consequently, the combination of these methods (PageRank and Topic Modelling) is a far more powerful approach for measuring user's influence [145].

Tweets, Retweets and Followers Kwak et al. [175] demonstrated that the number of followers is highly correlated to PageRank, whereas there is low correlation between followers and retweets and between PageRank and retweets. Surprisingly, Lerman et al. [182] reached a contrasting conclusion that there is a strong linear correlation between number of followers and retweets. The same study also showed that a good predictor for the number of followers is the number of followings, demonstrating that diversity of information in tweets is often rewarded by more retweets. Nevertheless, studies have shown that there is a strong correlation between number of followers and the number of different users that they usually retweet [82], thus when retweeting, users exhibit a strong favouritism towards certain users. The depth of the retweet pattern of a posted URL can give information for the significance of the URL and the influence of the user who first posted it [39].

In general, there is a positive correlation between number of followers and number of tweets (the more the followers, the higher the activity). It is indicative that in Twitter, 10% of users have 10 or lower followers and rarely tweet. At the other end, it is easy to spot "celebrities" by measuring the ratio of tweets versus followers. This is because the positive correlation between number of tweets and followers stops for users that have more than 10.000 followers. For these users, the high number of followers is due to their social status, rather than the quantity of their tweets [175].

Homophily

Homophily is the level at which people with common interests tend to associate in public environments, like social networks [210]. In order to measure this property we need

Table 3.1: The most common user properties related to the social graph: influence, centrality and homophily

Property	Meaning	Measurements
Influence	How important are users who follow me?	PageRank [175, 327] Hirsch Index [145]
Centrality	How centrally am I placed on the social graph?	Betweeness Centrality [106]
Homophily	How close am I to other users having the same interests as me?	Similar profile [175] TwitterRank [327] Lists [348]

a dataset enriched with user's activity (or interests), since the social network itself does not convey this information. For this reason, researchers are using either the text from the messages, or they are utilizing meta-information of the social graph provided by the social network service. In the first case, a common line of work is to perform a Topic Modeling analysis, in order to extract the different topics present in a set of messages. One of the most common methods for Topic Modeling analysis is the Latent Dirichlet Allocation (LDA) [48] [140]. After topic modeling, we can assess the degree on which a user is interested in a specific topic. Incidentally, this also measures the influence of this user on this topic. Finally, the correlation between the proximity of users in the social graph and the degree of shared interests gives an estimation of the homophily [327].

When utilizing meta-information from the social graph, homophily is measured according to the similarity of the time-zone, popularity (number of followers) [175], or the similarity of the subgraph in the vicinity of a user's node, like in the PageRank algorithm [57]. Another source for information regarding user's interests is the self placing of a user's followers into certain lists. The name of the list can give information of the primal identity of other users. By exploiting this information Wu et al. [348] revealed that there is a strong homophily, as expressed in retweets among celebrities, bloggers and media (in declining order of homophily).

On Table 3.1 we briefly present the meaning of influence, centrality and homophily along with the major studies that investigate these features.

3.2.1 The Social Graph as a whole

Graphs representing social networks have been a major research area long time before the era of OSNs. Freeman [118] has presented a thorough historical analysis that starts from the beginning of the 20th century. In this section, we present the main efforts to measure properties of the complete social graph of Twitter. A summary is available in Table 3.2.

Each measurement gives us intrinsic insight of the characteristics and dynamics of the network.

Degree of separations

Perhaps the most notorious property measurement of the social networks is the "degree of separations". This is the average number of "hops" that we need, in order to traverse the social graph from any random user to any other random user. This property has become famous even from the early ages of social studies, since in 1967 it was discovered that two random individuals are no more far apart than 6 hops in the social graph that represents real life (no virtual acquaintances) [220]. This finding was publicized as the "six degrees of separation" phenomenon or else the "small-world" phenomenon. Due to the easiness of which relationships can be created in social networks, this number is expected to be smaller.

This property is usually assessed by measuring the average length of all shortest paths of the network. This was measured to 4.12 on Twitter [175] and 4 on Facebook [37]. Another property, the diameter, is the longest of all possible shortest paths and portrays the linear size of the network. Since, this property is sensitive to distant outliers of the network, the 90th percentile is more commonly used, called the "effective diameter" [185], estimated in 2010 to be 4.8 [175] for Twitter.

Assortativity

Assortativity measures the correlation between properties of adjacent nodes. We can perceive assortativity as a method to measure the average homophily of the network, if we focus only on node properties (and not the content of users' posts). The most common type of this measurement is the degree assortativity. It has been suggested [157, 235] that assortativity can distinguish social networks from other "real life" networks. The rational is, that when a new edge (i.e. friendship) is formed in a social network, it tends to reach a node with similar attributes. This is not the case for other "real life" networks, like biologic or distribution networks, which tend to reach maximum entropy and their assortativity index is negative (also called disassortative networks). Indeed the assortativity of Facebook, at 2011, was 0.226 [316]. Kwak et al. [175] argues that since Twitter's social graph is directed, measuring assortativity is not feasible. Nevertheless Myers et al. [230] measured all four possible degree assortativity indexes (in- and out- degree of the source combined with the in- and out- degree of the target) and found 2 assortative and 2 disassortative.

Another way to bypass this limitation of measuring assortativity is to consider an undirected social graph, by taking only reciprocal connections (i.e. two users are following each other). This approach has been used to measure the assortativity of reply-networks [50].

Table 3.2: The most common properties of the social graph of OSNs and their respective measurements on Twitter.

Name	Measurement	Study
Degree of separation	Average hops between two random users 4.8	[175]
Distribution of node degree	In-degree is power law $\lambda = 1.35$ Out-degree is log-normal $\mu = 3.56$, $\sigma^2 = 2.87$	[230]
Average node degree	Average followers: 557.1 Average followings: 294.1	[39]

Assortativity	The number of my X is the same as the number of Y of the users that I follow:
	X = Friends, Y = Friends: 0.272
	X = Followers, Y = Friends: 0.241
	X = Friends, Y = Followers: -0.118
	X = Followers, Y = Followers: -0.296
		[230]

Reciprocity	Percentage of users that follow me back: 78%	[175]

Reciprocity

The social graph can also indicate the primal purpose for which a user uses a social network. This can be measured with the "reciprocity" value, which is the degree of which a user is followed by the users that she follows. A low reciprocity shows that the user is using the social network mainly as a source of information, whereas a high reciprocity shows that the social network is mainly used for communication among a users' peers. Reciprocity has been measured for several social networks like 68% for Flickr [71], 84% for Yahoo! [172] and 78% for Twitter [175].

3.3 Attacks and exploits

3.3.1 Spam

Estimating the performance of a spam defense system A well-designed spam defence mechanism should have two main characteristics: accuracy and efficiency. Accuracy is measured on both sensitivity and specificity. High sensitivity means that the system correctly identifies spam content in a high ratio. High specificity means that the system has a low number of misclassified non-spam content as spam (or else false positives). In general a non-spam tweet misclassified as spam should be more penalized than a misclassified spam as a non-spam. This is because flagging or even hiding legitimate content from the user may affect more the user's overall experience from the service, than dealing with a spam that eluded detection. Users are more accustomed to be exposed to spam than having legit content being hidden from them. A general consensus for acceptable sensitivity and specificity values are 80% and 99%, respectively [139, 242, 243]. Sensitivity is equal to 1 minus the true positive rate.

Efficiency measures the time overhead added to the system (or to user experience), by employing a specific defence mechanism. A very elaborate and complicated defence, regardless its success, might render a service useless, if it uses a considerable amount of time. Acceptable efficiency values are in the range of half of second. This includes both the amount of time taken to extract the features and to classify a given tweet [139].

Another consideration is that spammers are adapting quickly to avoid tracing mechanisms. Even if a defence is successful for current data, there is no evidence that the method can be robust on future spam deployed campaigns. This robustness is rarely discussed in existing studies [285].

Account hijacking Through account hijacking a single spammer can "own" thousands of legitimate accounts and orchestrate massive spam campaigns, unbeknownst to them [139]. Account hijacking in Twitter can happen either from brute force password guessing [347] or from phishing techniques [12, 144]. It is estimated that $520 millions were lost due to phishing attempts in 2011 [12]. A study from 2014 [302] revealed that out of 168 million users, 14 million had their accounts hijacked and 5 millions were deliberate fraudulent accounts. Hijacked accounts account for 69% of total spam in Twitter and the probability of users becoming victims is correlated with the number of victims that they follow. The authors also challenged one of the most profound beliefs regarding security in social media: "Only novice users can get hijacked". In contrast, they found that accounts that had many years of frequent online presence with hundreds of thousands of followers were victims as well. Social consequences are from abandoning an account (1 out of 5 victims), to losing online friends (1 out of 2 victims). This finding signifies the importance of introducing better spam defence mechanisms, as well as, to raise awareness of the public on this issue and urge users to be suspicious and to adopt basic practices for secure browsing.

Because of account hijacking, there is a crucial distinction between spam classification versus spammers identification [46]. Although spam content is very distinguishable, spam accounts can be in reality hijacked accounts that post a mix of legit and spam content. This was confirmed in a study of 100 million tweets at 2014 [17], in which the authors identified two very distinct patterns of spam accounts: The first had the same tweeting and social patterns to legit users, whereas the second had in average more followings and lower betweenness centrality. Therefore, a spam message in Twitter, does not necessarily means that it was sent from a dedicated spam account. For this reason, the accuracy of tweets classification methods (such as [46, 318]) are usually higher than methods for account classification (such as [20, 46, 115, 180, 318]).

Fake accounts Another popular technique is to simply create multiple accounts and have them exhibit a tweeting pattern that attracts a fair amount of followers [180]. After acquiring a critical mass of followers/targets these accounts can tweet spam content, along with harmless tweets. Of course Twitter has explicitly disallowed this practice and has built defences against it [311].

In 2018, Twitter announced [269] that it has improved its spam detection techniques and as a result it suspended 70 million accounts. Twitter also "challenges" 9.9 million accounts per week and has also posed actions against accounts that are offered as followers, in exchange of money. Twitter estimated that after these actions, the average number of followers of their users will drop by 4. For example, one particular company, Devumi, has sold 200 million Twitter followers, from a collection of 3.5 million fake accounts [86].

Link farming and Trend-jacking One of the main objectives of spammers is to augment their targeting audience, or else to increase their followers base. To achieve this, they usually engage in activities to make appear a spamming account as "interesting" and "informative". One of the most widely-used techniques is to re-post URLs to popular content such as news items, product releases and trending Internet memes. This practice is also known as "link farming" [128]. Other techniques to artificially increase influence are (1) adding mentions to popular users, (2) retweeting legitimate popular tweets and (3) adding hashtags with trending topics. The latter technique is called *trend-jacking* [203] and Grier et al. [139] revealed that 14% of trending topics are generated exclusively from spammers. The goal of this attack is to masquerade the spam message to make it seem innocuous and blend in with numerous other legitimate tweets about a specific topic. This technique also takes advantage of the very popular and efficient search functionality of Twitter. To put this in perspective, on 2018 Google served 3.5 billion searches daily [197], on 2016 Facebook's search engine received 2 billion queries per day [89] and on 2014, Twitter served 2 billion queries per day [229]. Although, these companies publish usage statistics sparsely, in a way that makes it difficult to compare between each other, it is evident that searches within Twitter constitutes a significant percentage of total searches for content on the web. Therefore, "hijacking" search results of Twitter, by mixing popular content with spam URLs, is a successful strategy.

3.3.2 Bots and the "fake news" epidemic

Fake accounts and automatic content posting can have more dark motives that simple financial gain, as it happens with spam. Today it is considered a cultural and social phenomenon the wide spread of "news" of questionable origin and validity. This phenomenon is called the "fake news" epidemic and is widespread on OSNs.

Prevalence of bots and main techniques for fake content circulation

In [300], the authors studied the infrastructure used to launch a massive misinformation campaign, in order to influence political conversations regarding the outcome of 2011 Russian's parliamentary elections. The attack was done from computers around the globe, consisting of 39% of blacklisted IPs, probably originated from compromised hosts. Of course, "opinion hijacking" through the use of Twitter bots is not only pertinent in politics. In [60], the authors revealed that Russian accounts that were active in US elections, were also spreading misinformation that promoted the anti-vaccination movement.

More recent studies revealed that bot activity is more wide-spread and more effective than it was thought. An analysis of 14 million tweets in 2018 demonstrated that a low number of bots (6% of total accounts) is enough to spread 31% of fake news [275]. This study also revealed two of the most successful bots' strategies. The first is to reproduce low-credibility content, as early as possible, (preferably less than 10 seconds) after the content is originally posted. This will give the chance, for the content, to be widely spread before it is refuted. The second is to target, through user-mentions, very popular users hoping that they will retweet and redistribute the content [286]. These techniques were called "automated amplification". An other study [317] that analyzed 126,000 stories, revealed that false news-items required, in average, 10 hours to reach 1,500 people, whereas valid news-item required 60 hours to reach the same amount of people.

Bot activities that target political campaigns is of special interest. On Twitter, hundreds of thousands of fake accounts seem to participate in orchestrated efforts to promote (or libel) certain political campaigns, an action that can be referred as "opinion hijacking". The phenomenon has been noticed during elections in countries like Russia [169], USA [66], Australia [324] and also in the Catalan referendum for independence [286].

Bots, Rumors and Fake news Detection

Given the sophistication of bot accounts, the task of bot identification is a very challenging task. The most well-known tool that employs machine learning methods for bot detection is botORnot [93]. This system has been expanded and renamed to Botometer. The Twitter Sybil Detector [19] (TSD) uses Machine Learning methods on 17 features and achieves a 95% detection ratio, although it fails to detect hybrid accounts (acting both as bots and as legit), which is a main drawback, given the hijacked nature of many accounts. TSD has made publicly available a Twitter Sybils corpus, that can be used for comparative analysis. DeBot [74] is a detection system that exploits the fact that bots tend to post content synchronously, in contrast to human. DARPA has challenged 6 research groups to perform bot detection for anti-vaccination campaigns [291]. One of the interesting parts of the challenge was that contestants had to distinguish anti-vaccination bots from other kinds of bots. Similarly, Chu et al. [80] tried to distinguish malevolent bots from bots that post benign content (called cyborgs, i.e. with automatic weather reports).

Another interesting line of work is to deliberately construct a variety of harmless bots, each with different "behaviour" and then study the number of followers or other influence metrics that they acquired [130, 217]. In a similar study it was found that a group of users did not find any differences on source credibility, communication competence and interactional intentions between tweets originating from humans or bots [107].

Bot detection is a different task than rumor detection. Although rumors and fake news exploit OSNs for rapid circulation, they do not have to be based on the existence of "bot armies". Most of the times, a well constructed rumor from a seemingly trustworthy source, regarding a recent and unexpected event, can be very easily propagated, even from experienced users. Consider that in 2013, a single tweet was enough for making the stock-market crash, for a short time [204]. Analysis of 18 features on a dataset between credible and non-credible tweets revealed small differences [242]. Zubiaga et al. [358] have studied and made freely available a dataset of 5.802 tweets, regarding 5 fatal events that sparked the circulation of many fake news (also called the PHEME dataset). Analysis of this dataset with deep neural networks yielded an accuracy of 82% [13]. Another valuable source of information for rumor detection is the retweet pattern of a tweet. When combined with simple linguistic analysis, it can identify trustworthy versus invalid information spread [214, 242]. An activity similar to rumor spreading, is *astroturfing*, which is the spread of positive comments, with the purpose of generating a fake "supportive" movement towards a person or a policy [259].

The defence against bots is to simply apply mechanisms for early detection and elimination. Shao et al. [275] estimated that by eliminating only 10% of bots is enough to significantly decrease their impact. Twitter itself applies a "quality filter" that removes possible automated content from a user's timeline [181]. Yet, regardless the automatic precautionary measures, the user's vigilance and well-constructed skepticism is a good defence against the fake news epidemic.

3.4 Sentiment Analysis

3.4.1 Analysis of languages other than English

The language is a major factor in sentiment analysis in Twitter, due to major differences on lexicons, syntax and semantics among languages. Today, sentiment analysis has been performed in all major languages like Spanish [24], Arabic [103] and Chinese [355], all of them following language specific techniques. What is perhaps more interesting are the multilingual techniques. In cases where, lexicons and text processing software is available for multiple languages, these can be combined to build multilingual systems [306]. Even more challenging are language agnostic techniques. On this area, a common approach is to use emoticons as inter-language sentiment indicators [91, 233]. Another approach is to employ machine translation, for example Google translate [40], to "normalize" tweets in a common language and then apply common sentiment analysis methods. When it comes to multilingual analysis in Twitter, it should be taken into consideration that, the demographics of Twitter users might differ significantly between countries.

3.4.2 Psychometric methods

Psychometrics [346] are the family of methods that attempt to assess various psychological traits of users, based on the activity and content of their online profiles. In cases where this research is focused on "happiness", the term "Hedonometrics" [101] is also used. The first study in this area [257] found statistical significant correlations between simple Twitter statistics (like number of followers) and "The Big Five personality traits" [44]. For example, the number of followers was strongly associated with the "extraversion" trait, the "imaginative" attribute was present only to popular users and the "organized" trait was present mainly to the influential users. The estimation of the "big five personality traits" through textual analysis was also associated with the choice of profile picture, through image analysis [191]. Another study found that happiness is assortative [50], meaning that proximal users (distance no more than 3 links) show correlation in happiness metrics.

Another line of work is to measure emotional variation. In [253], the authors applied sentiment analysis techniques in a corpus of 35 million English tweets, concluding that tweets with high emotional divergence get retweeted more often. Although the polarity of the tweets does not influence the probability of retweeting, the emotional divergence does have a measurable impact. Bollen et al. [52] measured the effect of important public events (like public holidays or general elections), on the collective sentiment. A study of 2018 [105], analyzed 800 million tweets from UK and measured the diurnal variation of 73 psychometric variables. The authors located two leading factors, named "Categorical Thinking" and "Existential Thinking", which peak at opposite time points during the 24-hour day. This study provided additional biological insights associating language use with the circadian rhythm. Emotional variation is measurable not only in the textual content of the posts, but also on the changes of the profile summaries and display names of Twitter users [328]. These changes were associated with the cultural self-identity of the users. Similarly, it is possible to measure the use of certain types of language between different cultures, an area of research that belongs to *linguistic relativity*. A study that analyzed 40 million tweets, measured the emotional variation of tweets between Canada and US, confirming the stereotype that, Canadians are in average more polite than Americans [282].

Identification of hate speech

Psychometric analysis is also used for the detection of hate speech in Twitter. Hate speech on OSNs is defined as online posts and commends, that are disgraceful towards individuals of certain race, religion, ethnic group or sexual orientation. Perhaps the largest available constructed corpora with hateful or abusive content in Twitter is from [117], which contains 80.000 annotated tweets. Available data for hate speech detection is also available from the comments section of Yahoo! [240] and Yahoo! Finance [100]. In [323], the authors define 11 criteria for hate speech identification and developed respective NLP techniques to quantify them. The study was performed on 16.000 tweets, manually chosen from an initial collection of 136.000 tweets. Another approach is to focus on specific events that can spark online debate, that might include hate speech. In a relevant study [64], the authors collected 450.000 tweets containing a hashtag related to a specific event and had 2.000 randomly chosen tweets, to be manually labelled as "containing hate speech" through the CrowdFlower service. Features included lexical and syntactic attributes, sentiment, user features like identity, number of followers and tweets features, like the presence of a URL. Another line of research is to identify hate speech that targets a specific group. For example, in [176] they built a classifier from 100 manually labelled tweets and evaluate it in a dataset that contained 24,582 tweets, with half of them targeting the black community and half having neutral content.

Hate speech has been a major issue in Twitter. A study from Amnesty International [21] reported that, in average, a woman receives an abusing tweet every 30 seconds and that women of colour are more likely to be targets of "troubling" tweets. Twitter has acknowledged this issue and acquired Smyte, a company that performs spam, abuse and fraud detection, with the sole purpose of addressing hate speech [315].

Health Monitoring

The users' timeline in Twitter reveals practical information, not only for the psychological, but also for their physiological state. NLP techniques on Twitter have been employed for tasks like monitoring of influenza epidemic [61], drug intake [196] and incidents of intestinal disease [357], obesity and diabetes [162].

Chapter 4
The average degree and the temporal growth rate of Twitter

In this chapter, we study the temporal growth rate of Twitter. Towards this goal, we delve into the existing models that describe the densification power law property of modern OSNs. We present a case study, where we apply the "Leskovec model" [184] on the average outdegree of Twitter. We assess this model in two samples of Twitter and study potential connections between events, during the early period of Twitter that may have affected its growth rates. We obtain these two graph datasets of Twitter through the available API and apply a heuristic, in order to approximate the graph link creation time, which is not available by the Twitter API.

Several background work has been done in order to measure the graph properties of many online social networks(OSNs). These studies' main target is to explain the behavior of their users and understand the growth dynamics of the underlying social graph regarding the dynamics that govern the creation and evolution of the network [290]. Background work shows that it is very common to measure the average node degree model and the node degree distribution of the network, as mentioned in chapter 2.

The degree of a node in an OSN is the sum of the *outdegree* and the *indegree* property. The *outdegree* is the number of edges with direction outward to the node, whereas *indegree* is the number of inward directed edges. Specifically, in Twitter the *outdegree* is defined as the number of followers of a user, while the *indegree* is the number of followings of a users.

4.1 The dataset and time snapshots generation

Acquiring a sufficient sample size for Twitter (according to [184], this sample size should reach approximately 90 million nodes) is prohibitive, given the current limitations of Twitter's API. We used two independent data sets for our analysis. The first dataset, consists of all the followers and friends of 92 million users which was obtained through the Twitter API, using the random walk network sampling algorithm. According to [184], random walk is the preferable method for capturing temporal graph patterns. This algorithm simulates a random walk on the graph by initially selecting a random user of Twitter and extracts all her friends and followers. Then, it randomly selects one of the newly added nodes and repeats this procedure. At every step, returns at the starting point, with a probability of $p = 0.15$, and begins a new walk. For time-efficiency reasons, we initiated 11 random walks, that were running concurrently. Each of these 11 "walks" had a different starting seed. We selected these seeds by randomly selecting 11 users, each one residing on a different geographic location (according to its latest tweet). These locations were: Canada, USA, Mexico, Argentina, United Kingdom, Greece, South Africa, Russia, Indonesia, Japan and Australia. Whenever a new sample (friend or follower of a node) was collected, we stored it in a Mongo database.

The same study that suggests Random Walk as an efficient sampling technique, [184], also addresses the issue of sufficient sample size for capturing graph metrics. A 15% sampling size is enough for measuring the graph properties of a graph as it grows and evolves. It is estimated, that, when the sampling happened (from September 2015 to April 2016), Twitter had 500 to 600 million users, who half of them were active users. Therefore, we estimate that 92 million users is a sufficient sampling size, since it constitutes 15% to 18% of the complete network. The average number of followers per user was 624 and the average number of friends was 763. We will refer to this as the *BIG* dataset.

The second dataset is presented in [175]. This dataset contains the entire graph of Twitter as of July 2009 and contains 40.8 million users. Additionally, we obtain from Twitter API all the followers and friends of these users, in the period from November 2014 to January 2015. Each user on this dataset has on average 210 friends and 214 followers. We will refer to this dataset as the *KWAK* dataset.

KWAK represents a sample of the early stage of Twitter, whereas *BIG* does not focus on a specific period. This allows us to focus on some interesting events that took place during the early period of Twitter. The evolution of the average outdegree is shown in figure 4.1 for the two datasets.

Considering the inactive users, they belong to one of the following sets:

- Users that have deleted their account, but are existent in KWAK dataset. In this case we want them to be included in the experiment since they were part of the graph, till the deletion. Deleted accounts in the second dataset cannot exist, since they are not available through the API.

- Users that are not logging in Twitter, but have an account . These accounts still have a number of followers that need to be in the graph. These consist part of the social graph.

- Users that are inactive and do not connect with other users. This set of users should also be included in the graph, for both datasets, since we will have nodes with very low or zero number of followers and will be depicted accordingly in the calculation of the average degree.

- Considering the spam account that have high number of followers and are later deleted this consists a second order effect. From the set of spammers that have high number of followers that do not follow the back, they are factors of connectivity for the graph , that should be also included in the measurement , as long as they exist in the graph (not suspended/deleted) and may affect the graph's connectivity.

The plot show a first validation of the "densification law" of the Leskovec model stating that the average degree increases over time. The average outdegree is increased over time which agrees with the "Leskovec model" and is evident in both datasets (*BIG* and *KWAK*). Also the average outdegree of *KWAK* peaks and drops after August 2009. Since all users in *KWAK* have subscribed in Twitter before that date, a fair proportion of them were inactive when the sampling happened (5 to 6 years later). No new followers are added to inactive users so the measurement of the average outdegree past that date (August 2009) with the *KWAK* dataset, is not representative of the real average outdegree value of Twitter. Nevertheless, this demonstrates that the influx of new users after 2009 compensated this effect and resulted in the increase of the average outdegree, as it is shown in the line representing the *BIG* dataset.

4.2 Generating time snapshots

We can apply a heuristic that produces a lower bound estimation of the creation time of the followings between the users, which is not provided by Twitter API. This heuristic is found in [212] and is based on the fact that Twitter's API returns the lists of followers and friends of a user, ordered according to the link creation time. This list contains the unique IDs of friends and followers and these IDs are increased monotonically. So, the order of these IDs also reveals the subscription order of these users: Between any two users, the one that has the lower ID, subscribed earlier in Twitter. Consequently, the link creation time of the following relationship, between users A and B, can be approximated by the most recent account creation time, among all users, that followed B prior to A. The computational complexity of this heuristic is $O(|E|)$.

The accuracy of the inferred link creation times of this heuristic, depends on the number of friends and followers of a user. The higher the number of friends, or followers of a user, the more accurate this heuristic is. For celebrities (users with more than 5000 followers), the link creation time is estimated with an accuracy level of several minutes but for users with lower number of followers or friends, the error can be higher reaching days or even weeks. Taking into account the range of time this study covers spans over 9 years

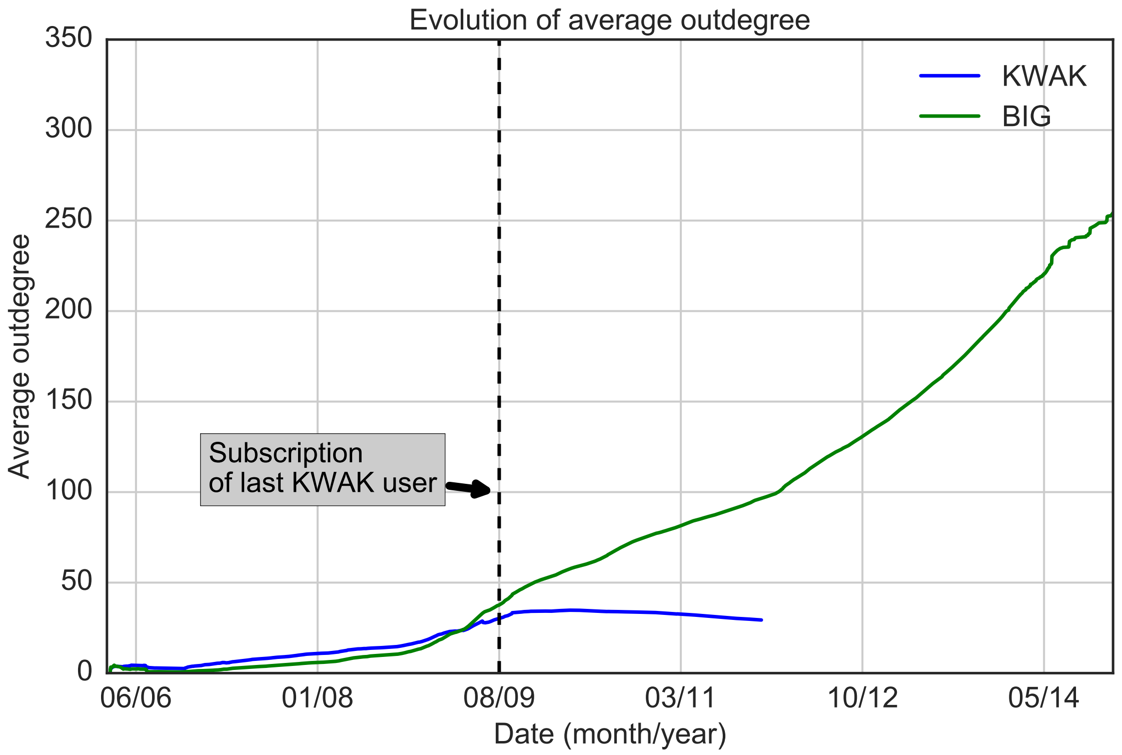


Figure 4.1: The average outdegree is increased over time which agrees with the "Leskovec model" and is evident in both datasets (*BIG* and *KWAK*)

(2006 to 2015), we do not expect these inaccuracies to introduce significant errors in our analysis.

When [212] was published, the heuristic assumed that users' IDs were ordered according to account creation time, or else that users' IDs were increased monotonically, but according to Twitter, this is not always the case. The strict monotonic order was guaranteed, until approximately 2011, after that, ID's kept increasing but the monotonic order is not guaranteed. In order to validate this, we plotted the Twitter IDs for 10 million random users, ordered according to the creation time of their accounts.

In Figure 4.2 we show that although the increase of Twitter IDs is not always monotonic after approximately 2013, it has a relatively canonical distribution. Twitter provided unique user IDs, on 10 million random users. The x axis shows the subscription date of users, and y axis shows the IDs provided by Twitter. In general, we notice a monotonic increase of these IDs over time. Nevertheless, there are small deviations, where the increase is not monotonic. In the subplot, we notice that, after a period, there are two different ID sequences. Our model, that estimates the friendship creation date, assumes strong mono-

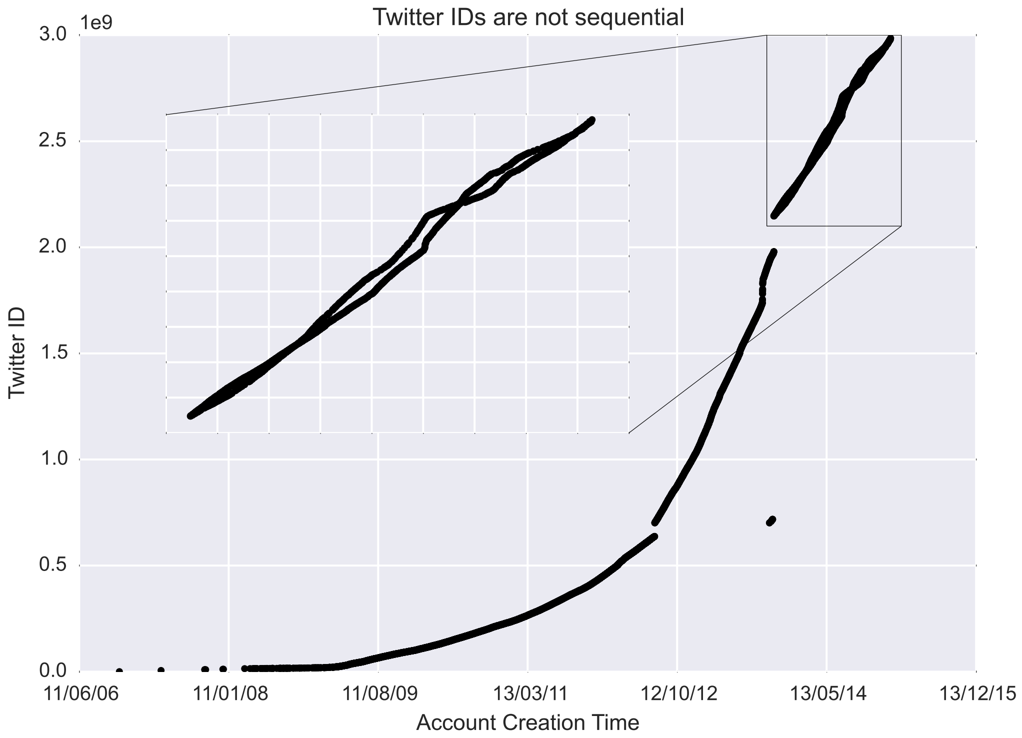


Figure 4.2: Twitter IDs of 10 million random users ordered by account creation time.

tonic ID increases (user A with ID greater than user's B ID is assumed to have subscribed later than B). Since we mainly focus on the early period of Twitter (before 2010), we do not expect these deviations to have any effect on our friendship date estimations. Additionally, since in the remaining of this thesis we focus on the early period of Twitter (before 2010), we do not expect this discrepancy to affect our findings.

4.3 The average outdegree of Twitter

The average outdegree and the diameter of the graph are the two of main properties characterizing the structure of OSNs.

The outdegree of a node (or else, a user) in Twitter is defined as the number of other users (or else "friends") that this user follows and the diameter is defined as the longest shortest path in the graph, among all pairs of nodes. Usually, measuring the evolution of an OSN over time, involves the study of the evolution of these parameters. In [14, 59], they propose that the main "laws" that characterize the evolution of OSNs are the constant average degree and the slowly growing diameter.

In an influential study, [186] they suggest that both these laws are wrong and fail to describe the evolution of many modern social networks and they propose an alternative set of laws, based on empirical observations: increasing average degree and decreasing diameter.

[186] also suggested a model describing the evolution of average degree. This model takes into account two parameters. The first is the Community Branching Factor, b that is defined as follows: if we model the graph as a tree of branching sub-communities, then b, is the fanout of this tree. Fanout is the maximum number of children that, a parent node might have in a tree. A large b is a characteristic of a dense network with tight communities. The second parameter is the Difficulty Constant, c representing the difficulty to create a cross-community link in the graph.

The growth of the average outdegree of an OSN depends on the relation between these parameters. Specifically, if the branching factor is higher than the difficulty factor, then the network's outdegree increases superlinearly. If these parameters are equal, it increases logarithmically, and if the difficulty parameter is greater than the branching parameter, the network has a constant average outdegree through time. In [186], the definition is given by the expected average outdegree of a network ($\overline{d}$) that is proportional to:

$$\begin{aligned} \overline{d} = &\quad n^{1-log_b(c)} \quad \text{if} \quad 1 \leqslant c < b \\ = &\quad log_b(N_t) \quad \text{if} \quad c = b \\ = &\quad \text{constant} \quad \text{if} \quad c > b \end{aligned}$$

while nodes (N_t) increase through time (t). In this equation, b is the community branching factor and c is the difficulty constant. In the case of superlinear growth (when $c < b$) the exponent $g = 1 - log_b(c)$ is a quantification of the growth of the network. We will refer to value g as the "Growth Exponent".

4.4 Fitting the model to Twitter

We are performing incremental measurements of the average outdegree, for every day of the dataset, in both datasets. Next, we fit the Leskovec model, to a "sliding window" of the average outdegree. The size of the window was 200 days and it was moved from the first 200 days of our dataset to the last 200 days, with a time-step of one day. We fit all three functions of the Leskovec model to the current window at each step, with the Levenberg-Marquardt algorithm [200]. Then we assigned the midpoint of the window to the model with the lowest root mean square error. The Levenberg-Marquardt algorithm also produced estimations for the b and c parameters.

In figure 4.3 we show the average outdegree of the graph, while the nodes are increasing in the *BIG* dataset. In the X axis we see the number of nodes at a given time and in

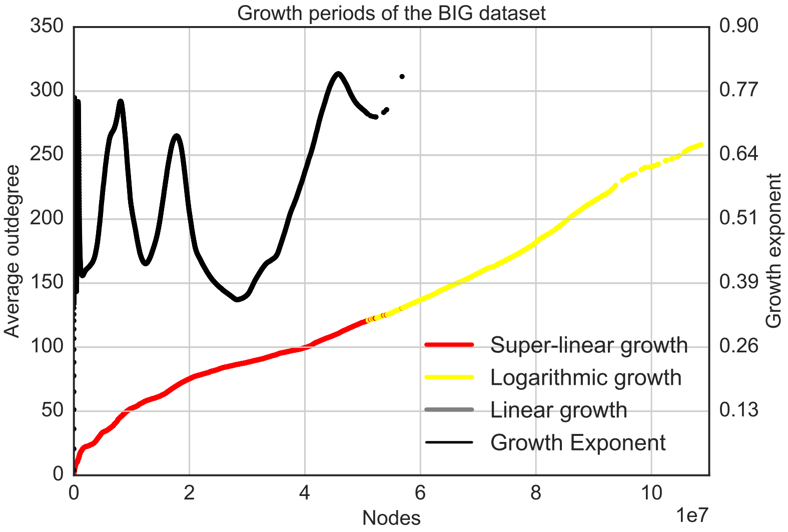


Figure 4.3: The evolution of the average outdegree for the *BIG* dataset, according to the nodes.

y axis we see the Growth Exponent of the graph (black line) and the Average Outdegree (coloured line). The red parts are following a superlinear growth, while the yellow parts follow a logarithmic scale; this pattern agrees with the Leskovec model. For the parts with superlinear growth, we also have plotted the the "Growth exponent", g. The plot of the "Growth exponent" is not available for the rest of the functions, since it is not available for the logarithmic and constant functions (yellow and grey colour, respectively). Initially, we notice that the Average Outdegree increases in a super-linear rate and after 50 million nodes it slows its rate into a logarithmic growth. Also, the Growth Exponent of the superlinear phase shows large variations that are indicative of various events that altered Twitter's evolution in its early period.

In [186] they predict that Twitter's social graph indeed, does not have a constant outdegree distribution. But the biggest part of the graph exhibits a superlinear growth. The growth factor during the superlinear growth, oscillates drastically reaching a maximum and then it is reduced. In the next section we present some events that may have caused these variations.

Figure 4.4 presents the same plot for the *KWAK* dataset. The last user subscription of *KWAK* happened when this dataset had 37 million nodes, so the large drop of the growth exponent after that is an artifact and not a real event. The small gap that is noticeable at the *KWAK* approximately at 06/09 nodes coincides with the death of Michael Jackson which caused a disruption in Twitter's online service. The plot, as in figure 4.3, illustrates the average outdegree coloured according to the function fitted by Levenberg-Marquardt algorithm (red for superlinear function, yellow for logarithmic and grey for constant). The plot of the "Growth exponent", g is only available for the red part of the plot, since it only available from the super-linear growth.

The *KWAK* dataset contains the users that have registered to Twitter anytime before the end of 2009. This means that the average outdegree measurement is not valid for the KWAK dataset for any day past 2009. The reason for this is that the users that the KWAK dataset contains are till August 2009. Specifically as mentioned in 4.1, the KWAK dataset contains: set (A) the complete list of user nodes (given by KWAK) till August 2009; and set (B) followers and friends (downloaded in 2015) . All the new user nodes from 2009 - 2015 are not existent in the graph, but we have the followers and friends for set (A) . We have the connections that these users (A) were creating till 2015, but not the new users added to the graph. Consequently, we cannot make any estimations of the overall outdegree distribution of the network, for any day past the end of 2009, based on these users, because followings in the network are happening at a higher rate from recently subscribed users, compared to older ones. So the outdegree distribution drops after the end of 2009, which does not reflect a real tendency for the network. The same applies for the growth exponent, we cannot produce reliable estimations when the window goes over the end of 2009. Nevertheless, we have a better resolution of the growth rate and the average outdegree of the graph, for the period marked from the start of Twitter until the end of 2009.

In figures 4.5 and 4.6, we show the temporal evolution of the average outdegree for both datasets respectively. The semantics of the lines in 4.6 are the same as in figure 4.3. Also as indicated in the aforementioned figures, the plot of the "Growth exponent", g is only available for the red part of the plot, since it only available from the super-linear growth. The only difference is that, x axis shows the date, and y axis shows the growth exponent of the graph (black line) and the average outdegree (coloured line) of the graph at that date.

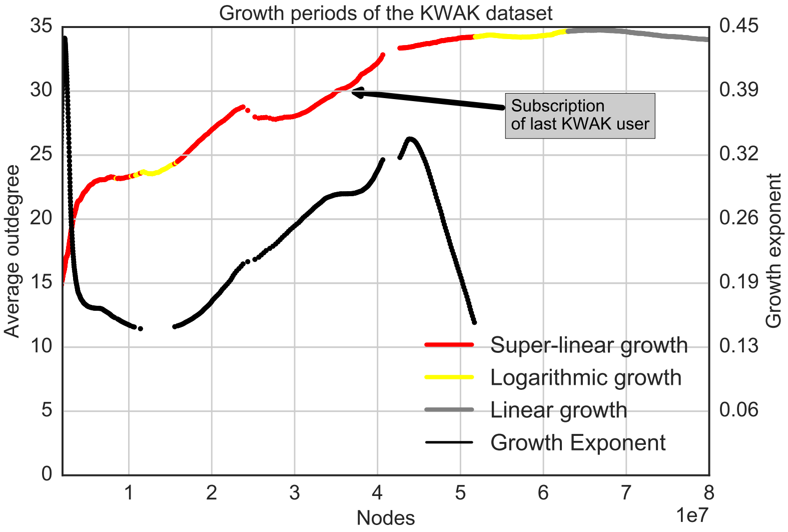


Figure 4.4: This is the evolution of the average outdegree for the *KWAK* dataset. The lines follow the same semantics as in figure 4.3.

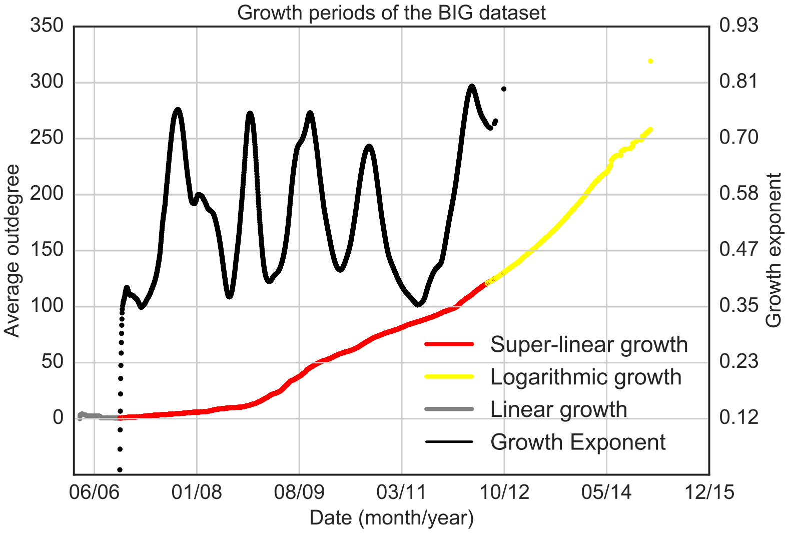


Figure 4.5: Here we plot the temporal evolution of the average outdegree and growth exponent for the *BIG* dataset.

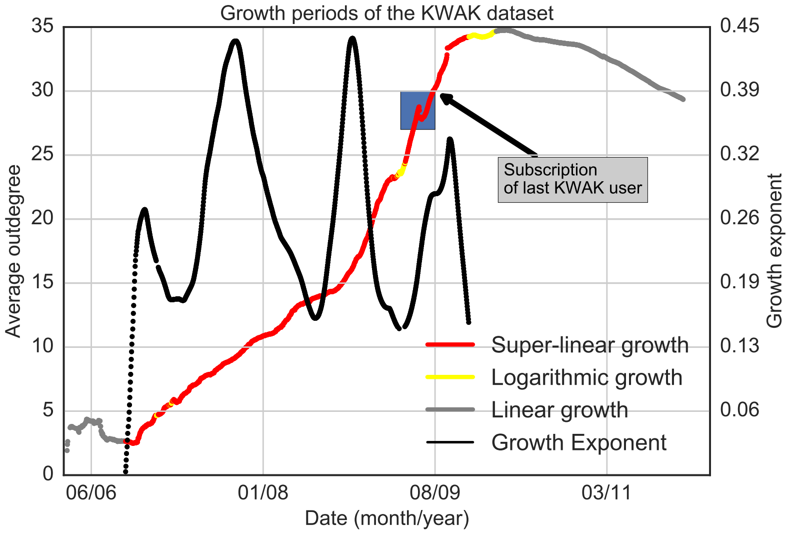


Figure 4.6: This plots shows the temporal evolution of the average outdegree for the *KWAK* dataset).

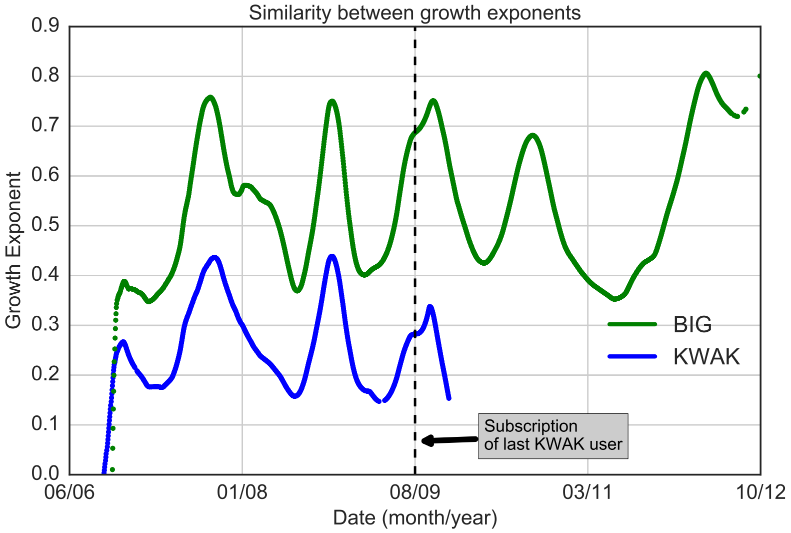


Figure 4.7: This plots shows the comparison of the Growth Exponent between the *BIG* and the *KWAK* datasets.

Finally, in figure 4.7 we plot only the growth exponent on the same time scale for both datasets.

The estimation of the Growth Exponent (g), depends on the number N of nodes in the graph ($\bar{d} = N^g$). The average outdegree ($\bar{d}$) between the two datasets, for the period before the subscription of the last *KWAK* user, is approximately the same (see figure 4.1). However, since *KWAK* focuses exclusively on this time period, it contains more samples (nodes) than *BIG*. So our fitting model algorithm generated higher g values for the *BIG* dataset. Nevertheless, we notice that despite the inherent differences between the two datasets (size, users), the fluctuations of the growth exponent delineate almost the same time periods. We see that the growth exponent can delineate various periods of increased or decreased superlinear growth. This can help pinpoint various time points, where the growth rate of the network was influenced by potential events might have taken place.

Figure 4.8 shows the first gradient of the growth exponent for both graphs. This figure shows clearly that although in different scale, the two plots increase or decrease with the same gradient. The estimation of the Growth Exponent (g) is sensitive to various parameters of the sampling method. Yet, the rate of increase or decrease of g, shows a relevant tolerance to these parameters. To validate this, we plot here the first gradient of g, between the *BIG*, and *KWAK* datasets.

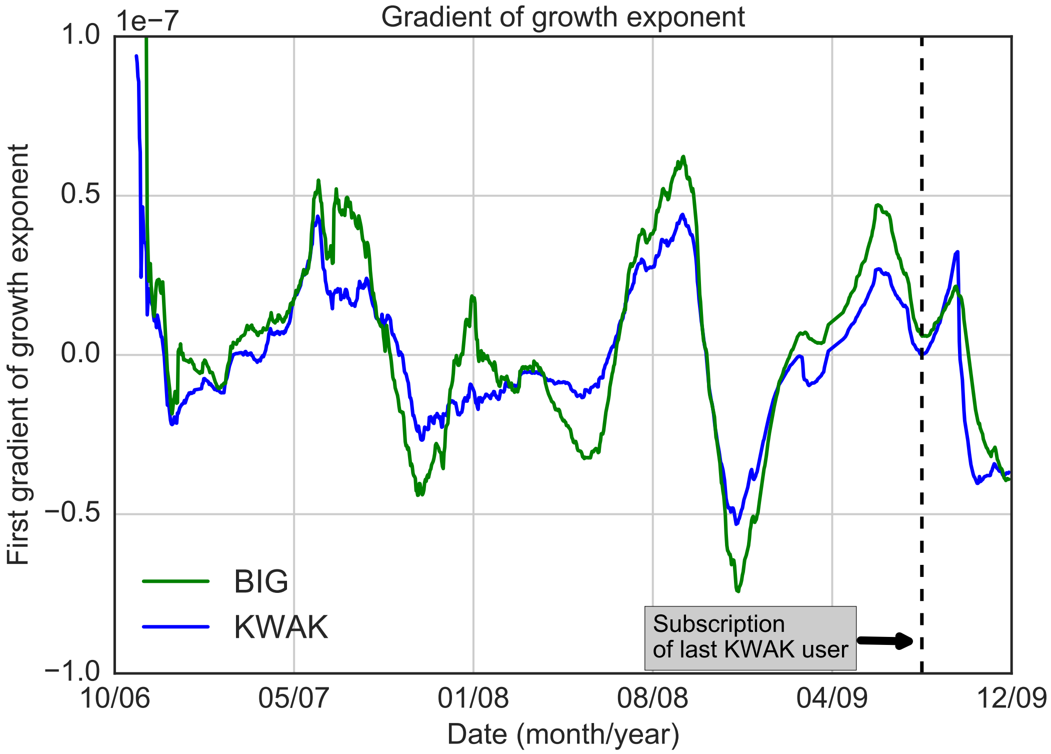


Figure 4.8: The first gradient of the growth exponent for both graphs.

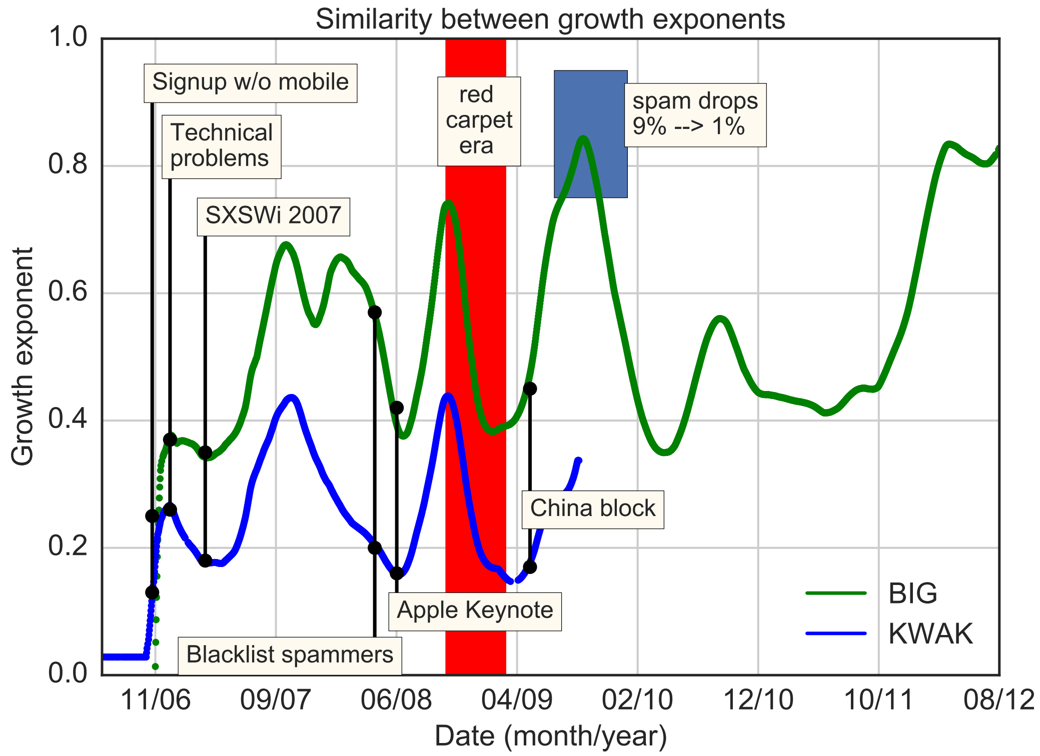


Figure 4.9: The events that may have influenced the growth exponent g of Twitter during its early period.

4.5 Events at the early stage of Twitter

As mentioned above, the growth exponent is able to capture various periods at the early stage of Twitter. These periods, are marked either with increased, or decreased superlinear growth. Figure 4.9 shows the growth exponent annotated with events that have influenced Twitter, according to Wikipedia [345]. When g decreases, then the average outdegree grows at a smaller rate. The periods of decreased g, indicate higher rates of addition of new users and periods of increased g indicate higher rates of new connections (followings). It is important to note that, in this study, we do not infer causal relationships between these events and the growth exponent, since a simple coincidence is not enough to justify a causal relation between an event and a growth change. The purpose of this analysis is to put these events into perspective, according to the changes of the growth exponent. Additional work is required, in order to quantify how these events might have actually affected (or not) the growth of Twitter.

However, the importance of some of these events (like the SXSW conference), has been validated by Twitter's officials. In the beginning (July 2006), Twitter was an experimental service developed exclusively for use with mobile phones. In October 2006 you could sign up without the usage of mobile phone [329]. This change marked a transition to a regular OSN where we notice a first increase in growth exponent. Several technical problems indicative of the service immaturity [6] at the end of 2006, may have slowed down Twitter's growth. Additionally, several rival services (e.g. FriendFeed, Pownce, Jaiku, Brightkite) make their appearance to attract potential new users [7]. The decisive breakthrough of Twitter happened in March 2007, at the SXSW conference [9], where Twitter won the top award and got a lot of attention. The user base of Twitter grew significantly, during this period.

Twitter applies its first action against spam in May 2008, by massive deletion many spam accounts. Whether spam increases or decreases the superlinear growth, is an open question, that we discuss below. In June 2008, a lot of blogs and websites were expecting that Twitter will not withstand the extreme traffic from Apple's keynote conference. However, Twitter did not have any failures, which was a sign of a transition to a more mature and stable service and as a consequence an increase of the growth exponent. The attraction of many personalities from the show business industry characterized the period from November 2008 to April 2009 as the "red carpet era" of Twitter. According to [159] 54% of the most popular Twitter users, started using Twitter during this period. The increase of growth rate is visible in the *BIG* dataset, but not in *KWAK*, and this is the only difference between the growth rate of the two datasets. We also measure the average outdegree per isolated day where we count the average degree of the graph for each day, without taking into account any previously formed edges, giving an evaluation of the density of the graph that was generated each day. We were surprised to find that this average degree was increasing each day until June of 2009, where it peaked and then started to decrease. We located two events that happened in this period. The first was the blocking of Twitter in China, and the second was the death of the famous pop artist Michael Jackson. In figure 4.11, we show these measurements annotated with these two events. On the following subsections we discuss how these events might have affected Twitter.

4.5.1 Blocking from China

Twitter was blocked from China blocked in early June 2009. We speculate that this might have reduced the average degree per isolated day, although we don't see any change in the growth exponent, as we shown in figure 4.11 To test this hypothesis, we search a dataset containing the user objects of 250 million users.

The user object is a data structure containing several meta-information about the user's profile, like the language, the location, the creation time and other profile preferences. The user objects can be requested from Twitter's API and they include the last tweet of the user. In this dataset, we look for users whose last tweet was tagged with geo-location information and we measure the percentage of those located in China. We also measure this percentage per year, according to the account creation time and according to the time this last tweet was posted. Unfortunately, Twitter, enabled geo-tagging of tweets in August of 2009 and it was very slowly adopted by users due to lack of support from Twitter clients [62]. As an effect, prior to 2010 the statistics based on geo-tagged tweets are not available. To tackle this issue we also watch the "location" field of user objects but this is a user-defined location field, so there is no guarantee that the actual location of the tweet is in China. Nevertheless, since it was the only location information available for tweets prior to blockage, we also measure the percentage of users that specified their location as "China" (in English or in Chinese) in their profile.

Figure 4.10 shows the percentages of Chinese users in Twitter, for each year, between 2006 and 2015. We use four different criteria to decide whether a user is Chinese or not.

- The black bars is, the account creation year of the users whose last tweet was geo-tagged in a location within China,
- The gray bars is the year of the last tweet for these users.
- The red bars is the account creation years of the users who are self-described as Chinese in their profile and
- The forth pink bars is the year of the last tweet for these years.

Since geo-tagging was enabled in August of 2009, we do not have geo-tagged tweets before that period.

This figure shows the percentage of Chinese users peaked at 2007. After 2007, the percentage drops even more than before the blockage takes place. From these measurements, the maximum decrease was 1.4% (from 1.6% at 2007 to 0.2% at 2009) for the account creation year of users that posted geo-tagged tweets from China. We believe that this large change cannot be attributed to the sudden blockage of only 1.4% of Twitter users, although in figure 4.11, we notice that in July 2009 the average degree per day starts to decrease.

4.5.2 Death of Michael Jackson

The excessive online traffic that sparked from the death of the famous pop artist Michael Jackson on 25th of June 2009, created an unexpected disruption of many websites including Twitter. In figure 4.6 the blue rectangle indicates a "bump" on the plot of average out-degree. This coincides with the disruption of the Twitter service from the death of Michael Jackson. From this technical problem Twitter recovered quickly. Researchers have used this event in order to study the propagation patterns of Twitter [351], as well as the emotional content of related posts [163]. This event might also have contributed to the in-

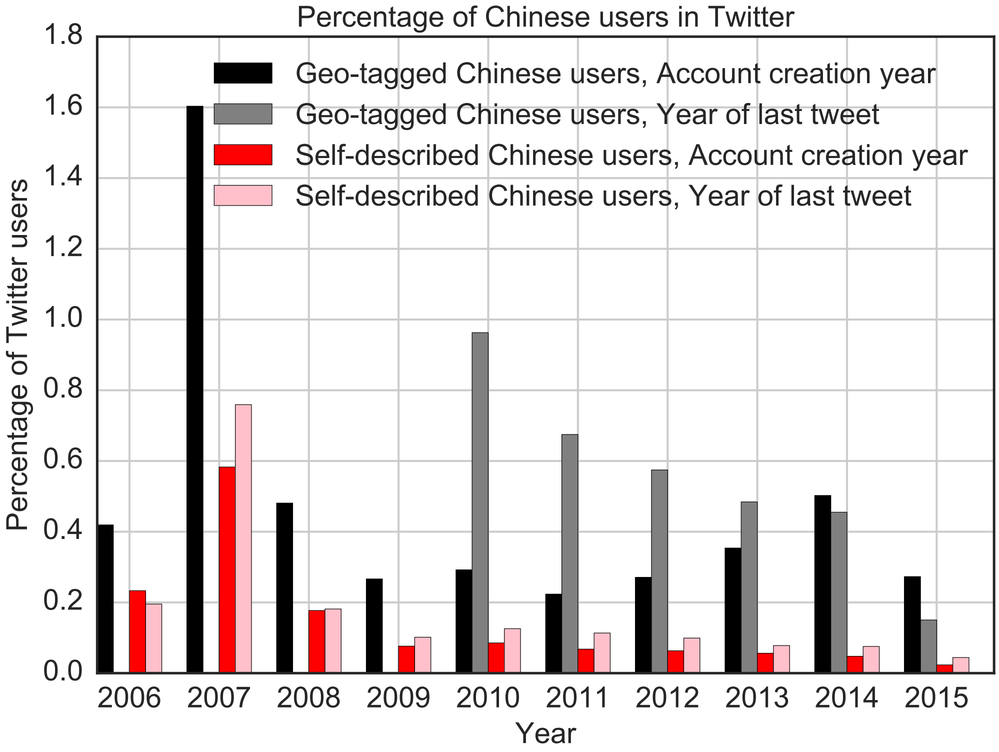


Figure 4.10: The percentages of the Chinese users in Twitter between 2006 and 2015.

crease of popularity of Twitter, in the long term, due to the publicity that this disruption reached.

In figure 4.11 we notice that before this disruption, the average outdegree of the daily graph was increasing linearly in time. On the day that coincides with the death of Michael Jackson, this increase stops abruptly. After that, the daily average outdegree decreases constantly and converges to a value close to 2. One hypothesis is that the death of Michael Jackson made Twitter suddenly increasingly popular, attracting users that enrolled in a high rate. Since new users have lower outdegree compared to older ones, this might have contributed to the decrease of the overall average. Specifically, for each day of the period: 1/6/2008 - 31/12/2010, we extract the following relationships, that happened that day on the *BIG* dataset. Next, we construct the social graph of each day and we measure its average outdegree, shown in the black lines. For each one of this daily graph, we also plot the number of nodes (blue) and edges (red). Notice that the daily average degree peaked between the dates when China blocked Twitter and the death of popular singer Michael Jackson. Also we see that the number of nodes and edges of the daily graph were stabilized on that period.

The explanation that we provide is purely suggestive. Further investigation of the cor-

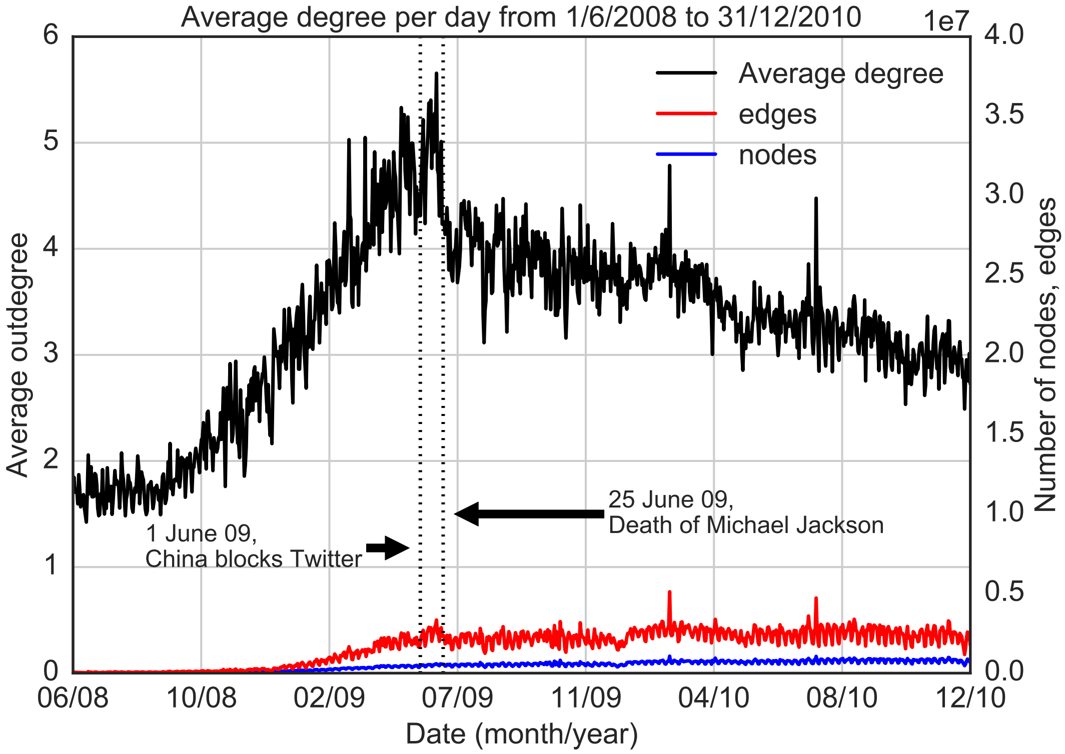


Figure 4.11: The average outdegree of the daily graph was increasing linearly in time until the disrupt.

relation of the average outdegree and this event would require access to a large part of the tweets of this period. Unfortunately this is not possible since Twitter does not provide access to data older than 10 days.

4.5.3 Spam filtering

Spam in Twitter has been an important issue [46]. According to related literature, the click-through rate of Twitter spam is significantly higher than mail spam [138]. In August 2009, the percentage of spam tweets in Twitter had reached the percentage of 9%, affecting its public image as a "clean" service that did not propagate spam or malicious sites. In August 2009 Twitter, in an effort to mitigate this, embedded a spam filtering mechanism on its URL shortening service. A report from Twitter points out that this technique reduced the spam percentage to 1% in February 2010 [8]. This period coincides with the beginning of a long decrease of the superlinear growth rate as shown as a blue shade in figure 4.9. The hypothesis in this case, is that, the superlinear growth rate was affected by spam accounts. Spam accounts in order to increase their target base, were following as many users as possible with the hope that these users would follow back, thus making them potential targets for spam or malicious URLs. This might have contributed to the increase in the growth of average outdegree. The application of the spam filter stopped or slowed down this promotion technique and stabilized the average outdegree.

Chapter 5
Trends and spam campaigns in Twitter

Users in Twitter can add a hash symbol (#) and a specific keyword, known in Twitter as a hashtag in order to enrich the semantics and assign context to the entire tweet by associating it with a topic. As mentioned in chapter 2, hashtags evolved to a social phenomenon. In Twitter popular hashtags and search queries are referred as the popular trends, trending topics or, simply, *trends* in Twitter.

Unfortunately, trends consist a very effective way for luring users into visiting malicious or spam websites, a technique called *trend-jacking* [203]. The purpose is to masquerade the spam message to make it seem innocuous and blend in with numerous other legitimate tweets about a specific topic, so the attackers collect information regarding the most popular (*trending*) topics and include them in tweets pointing to spam sites. By this way they increase the readability of spam tweets as they have a higher possibility of reaching large audiences through the search function of Twitter. Another approach is to collect legitimate tweets and add URLs (or replace any already contained) pointing to spam sites. The URLs are shortened in order to meet the restriction of the tweet size of messages, which also has a negative side-effect: the website that the URL points to is hidden and the user only sees the address of the shortening service along with a random identifier.

This chapter demonstrates a comprehensive analysis of this phenomenon. Initially, we obtain the dataset of the tweets that contain popular trends through Twitter's API and then we extract the contained URLs and measure the number of contained trends as well as other features. We use 86 different Real-time Blackhole Lists (RBLs) to get spam status of these URLs and the *ground truth*, i.e., spammers within our dataset. Then we build classifier to detect spam tweets, by leveraging information regarding specific features of the Twitter ecosystem which differentiate legitimate from malicious users. Also, this analysis detects a new masquerading technique that spammers use in Twitter.

Spam campaigns are centrally orchestrated mass efforts to distribute large amounts of tweets. The origin of these messages are mainly hijacked accounts [124] that are promoting a specific service or product. This analysis detects a specific type of spam campaigns that evade detection by masquerading URLs as Google search results and follow a conservative approach to propagating spam. Specifically, these type of campaigns try to attract victims by offering to increase the number of accounts that follow the user. We will refer to this class of spam as *GMF* (Get More Followers). The users are being tricked to give permission to a malicious website, e.g., by trusting a site that advertises some type of meta-analysis of the user's account. These fraudulent services most of the time offer usage statistics reports, analysis on user's followers, which followers retweet most of the user's tweets or who viewed the user's profile. The interesting part is that most of the services these sites offer can be actually performed without giving any special permission to the third-party. Even more surprising is the fact that there are legitimate sites that perform most of this analysis, for example: [116].

This points to the fact that there is a serious lack of public awareness regarding the actions that third-party services can perform when explicit authorization is permitted, as opposed to what actions can be performed by simply accessing a user's public data. These "get-more-followers" campaigns [289] are a serious threat since it is a multi-million dollar scheme [251], either as a paid service or a free one.

Evidently, attackers can acquire access to fraudulent Twitter accounts in underground markets [303]. Thomas et al. [301] analyses this phenomenon by presenting the techniques and attributes of spam campaigns from 1.1 million accounts suspended by Twitter. Adversaries in these campaigns follow a stealthier approach compared to other spammers, as they manage to masquerade the malicious URLs behind a legitimate popular site as Google. Spammers also try to maximize the coverage of potential victims by employing a much larger number of trends compared to other campaigns. We analyse all the available Twitter features and we show that the amount of different trends, included by a user in tweets, exhibits the highest divergence between spammers and legitimate users.

We implement a classifier based on the findings mentioned above that separates spammers and legit users. Specifically, the classifier on "Get More Followers" class spam, achieved a True Positive Rate (TPR) of 75%, which is comparable to existing studies, while maintaining a False Positive Rate (FPR) of 0.26% that is significantly lower than existing studies. We also extend the classifier to focus on individual tweets rather than users with similar results (81% TPR and 0.58% FPR). This classifier is computationally very efficient since it takes advantage of Twitter-provided features that require minimal pre-processing. The final step of this analysis is the visualization of the GMF campaigns that reveals that the spam domains were posted by thousands of users but originated from only 2 IPs. This analysis reveals that GMF spammers are most probably regular users with exploited accounts that interject the owner's legitimate tweets with certain spam tweets while RBL spammers have a higher probability of being dedicated spam accounts. Also a spammer is located by 1.9 RBLs on average, in a single day. Additionally, we show that trending topics are being exploited in the first day of their appearance and approximately half of the trending topics have been used at least in one spam tweet. Finally, we reveal link farming [128] by

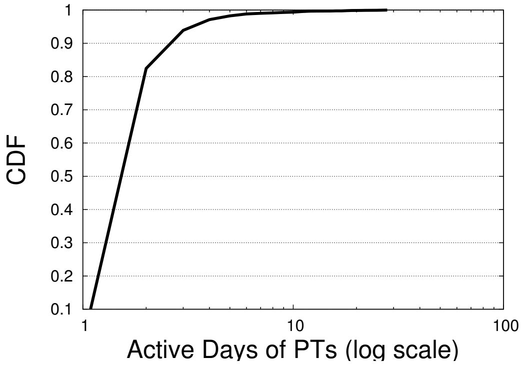


Figure 5.1: Prevalence of PTs for March of 2014.

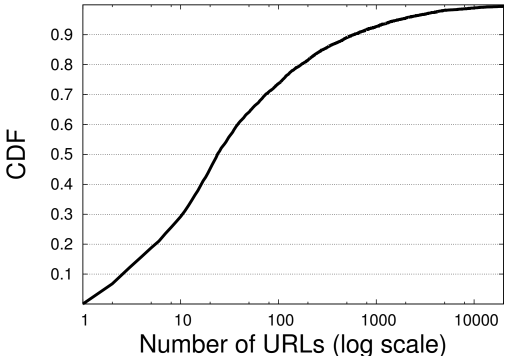


Figure 5.2: Number of different URLs associated with a specific trending topic.

graphical representation of the "Gain More Follower Campaigns".

5.1 Methodology

We retrieved from the public API available from Twitter the daily Popular Trends (PTs) and a subset of tweets containing these trend. We implemented a URL expansion method that allows us to collect the final URL from the shortened URLs contained in these tweets. On average, we downloaded 240 PTs and 1.5 million tweets, per day. Within a period of three months from January till March of 2014 we collected 150 million tweets.

In Figure 5.1 we show the Cumulative Distribution Function (CDF) of the duration of popular trends during March of 2014, where we notice the ratio of trends that were active for less or equal days than the values in x axis. We see that 80% of the trends are active for 2 days or less, although certain trends remain active and very popular for more than 20 days. These trends are usually more generic or associated with a specific geographic region. An example of long-lasting trends are America, Florida, Russia, Starbucks, Netflix and Disney. We also notice the diverse number of URLs posted per trend. In Figure 5.2 we see the CDF of the number of different URLs associated with a particular popular topic. We notice that approximately 90% of trends are associated with less than 1,000 URLs.

5.1.1 Feature Extraction

The next step was to extract the set of diverse user metrics consisting the *features* of our classifier. For each user we extract: The total number of tweets, the number of Total and Unique Popular Trends, Hashtags, User Mentions and URLs, the number of followers and the number of followings.

Groups of potential spam campaigns.

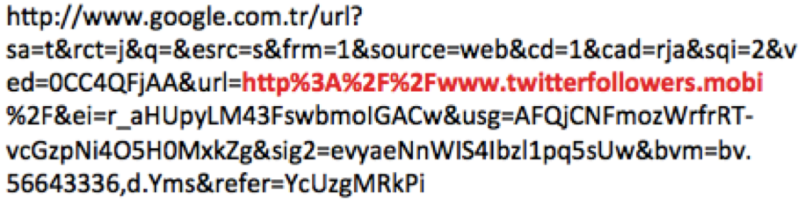
The next step after having obtained the metrics above, is to build a graph containing all users and group them according to the URLs they post, i.e., creating subgraphs for each URL, that contain all the users that included it in a tweet. We remove all link nodes that have a degree smaller than 10 and extracted the domain names of the remaining URLs, in order to keep the URLs that are posted in bulk.

This process result in 24.000 different domain names during our 3 months collection period. Next, we need to set the ground truth by recognizing the spam domains between our collected URLs and we have two methods for that: (1) Initially use various blacklists (2) Apply a heuristic that identifies spam belonging to the GMF class.

5.1.2 Extracting ground truth

In this study we distinguish spam in two classes. The first class of spam is called RBL and is defined as tweets that contain URLs that have been blacklisted by Real-Time Blackhole Lists. The second is called GMF and contains URLs that belong to the "Get More Follower campaigns". The RBL list contains 1820 domains. These are 1911 domains identified by RBLs minus 91 domains that were identified as not spam (manually) and were removed from the RBL lists. The Get More Followers domain blacklist was composed manually with 106 domains. So our ground truth was consisted from the spam RBL set, which are the tweets that contain RBL blacklisted URLs and the spam GMF, which are the tweets that contain the GMF blacklisted URLs. Below we explain each of these classes.

Real-Time Blackhole Lists (RBL) Initially, we query various Real-Time Blackhole Lists (RBLs) to recognize spam campaigns contained in our dataset. An RBL is defined as a list of IP addresses published through the Internet Domain Name System (DNS), offering the ability to query a domain name in real-time. These lists are most often used by mail server software and publish addresses linked to spamming behavior. Although these lists offer a fast and low-bandwidth method for spam detections, they have the disadvantage that they exhibit a lag time for updating and including new spam domains [139]. This disadvantage is more severe in Twitter because domains contained in tweets reach a wider range of users



```
http://www.google.com.tr/url?
sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&cad=rja&sqi=2&v
ed=0CC4QFjAA&url=http%3A%2F%2Fwww.twitterfollowers.mobi
%2F&ei=r_aHUpyLM43FswbmolGACw&usg=AFQjCNFmozWrfrRT-
vcGzpNi4O5H0MxkZg&sig2=evyaeNnWIS4Ibzl1pq5sUw&bvm=bv.
56643336,d.Yms&refer=YcUzgMRkPi
```

Figure 5.3: An example of a get-more-followers spam link obfuscated as a Google search result.

much faster than in email.

This is because while mail is checked in arbitrary time points, content in Twitter is accessed almost instantaneous. This makes delay in spam flagging in Twitter very crucial. Nevertheless most spam identification techniques use RBLs as a first step in identifying spam. Here we use 86 RBLs to get an initial set of spammers that will be used for extracting features for spam classification.

Get-more-followers (GMF) campaigns We perform a manual inspection of the collected URLs to identify obfuscated spam links as Google search results, during which we observed a large percentage of malicious links. A user potentially embeds a Google search results URLs in a tweet by just coping URLs from a Google search results' page and paste them into a message.

Although the final domain is not Google, when copied they actually are Google URLs. Initially, on click, Google redirects the user to the desired link but and then records the fact that this specific user clicked this link. A detailed inspection showed that spammers exploit this and use it as a mechanism for *link obfuscation*. An example is presented in Figure 5.3.

So in this way, spammers can obfuscate spam URLs as Google results and conveniently bypass any blacklists or filtering mechanisms of Twitter. We extract all Google results URLs from our dataset and recognized 44 domains that belong to the get-more-followers (GMF) domains. This list was appended with another 62 domains containing the word "follow" that we searched in RBL blacklists. The total 106 domains discovered mapped to 33 different IP addresses.

5.2 Data Analysis

Next, we follow the methodology presented in Section 5.1 in order to proceed to the data analysis of the collected dataset. We flag as spam 1,911 domains from our initial dataset, which is the 7.9%. In order to do this we utilize the RBL lists mentioned above and the "get-more-followers" heuristic. Note that the 1,911 domains traced back to 1,429 different IP addresses. Additionally, we took out 91 domains that were obvious false positives which we recognized through manual inspection. Another very interesting point is that from the 4,593,229 different URLs contained our dataset, the 250,957 of them pointed to a single spam domain. That makes the 5.4% of all URLs, which is significantly higher that the 1% that Twitter reports [78]. This shows that a remarkable amount of spam can bypass Twitter's spam detection mechanism.

The next step after we obtained an important labelled dataset of spam URLS, was the identification of all spam users by examining which users contained spam URLs in their tweets. The result of this procedure was 590,000 out of 8.2 million total users who have posted at least one spam link in their tweets. This accounts for 7.2% of the users. These users are most likely compromised accounts of victim users rather than spammers, as mentioned in chapter 3.

The following step was to check how Twitter features were differentiated between legitimate users and spammers (including compromised accounts). We present the values of the collected features measured per day divided with the number of active days of a user. The number of active days for a user is defined as the total number of days this user has posted at least one tweet, that was collected by our system. We provide a simple example for reasons of illustration. Suppose we have a user who sent the trends #T1, #T2 and #T3 within one day and the trends #T1 and #T4 on another. The specific user has 3 Different Trends (DT) for the first and 2 for the second day. Since the number of active days for this user is 2, we calculate the average value of DT as 2.5. The weighting is essential in order to produce metrics that are independent of the measurement period.

The results indicate that all metrics exhibit a larger mean value for spam users compared to legitimate ones. From these features, total trends and total hashtags exhibit the highest mean increase, (2-fold). We also notice that the different trends and the different hashtags exhibit similar average increases (both from 1.0 to 1.6) which is expected given that these two features (PTs and Hashtags) are similar. Additionally, in the case of the GMF domain dataset the difference of the average values for Twitter features between spammers and legitimate users is wider. Surprisingly, we don't observe any significant difference for the number of User Mentions compared to the RBL+GMF dataset. This leads us to the conclusion that this feature is not exploited by spammers.

Table 5.1 shows the average values of all metrics when applying both RBL and GMF detection techniques and when applying solely the GMF campaigns detection method, where we have the RBL+GMF and only the GMF domains. UM is User Mentions, metrics regarding total numbers are denoted by T, whereas unique numbers are denoted by (U).

Notice the difference of the average values for Twitter features between spammers and legitimate users is wider in the case of the GMF domain dataset. We can conclude that GMF spammers exploit Twitter features in a more prominent fashion. Specifically, we see

Campaigns	**RBL + GMF**				**GMF**			
	Mean		**p95**		**Mean**		**p95**	
	S	**L**	**S**	**L**	**S**	**L**	**S**	**L**
Tweets	2.5	1.6	8.0	4.0	1.7	1.6	4.5	4.0
Active Days	3.9	1.6	**16.0**	4.0	1.5	1.8	4.0	5.0
Trends (T)	2.6	1.3	**10.0**	4.0	5.0	1.3	**16.0**	4.0
Trends (U)	1.6	0.9	6.0	3.0	3.8	0.9	**10.0**	3.0
UM (T)	1.8	1.4	7.1	4.3	0.6	1.4	4.0	4.6
UM (U)	1.4	1.2	5.2	3.9	0.4	1.2	3.0	4.0
HashTags (T)	2.6	1.3	**10.1**	5.2	4.4	1.3	**13.5**	**5.6**
HashTags (U)	1.6	1.0	5.5	4.0	3.4	1.0	**8.0**	**4.0**
URLs (T)	1.6	1.1	4.0	2.0	1.4	1.1	3.0	2.0
Spam URLs	1.6	NA	4.0	NA	1.4	NA	3.0	NA
Blacklist hits	1.9	NA	5.0	NA	NA	NA	NA	NA

Table 5.1: Mean and 95th percentile values of features for spammers (S) and legitimate (L) users for both two studies.

that total trends and total hashtags exhibit a 4-fold increase for spammers. This observation is backed from the fact that user mentions from spam users are lower than from legitimate users in the GMF experiment, given that the GMF is a more spam targeted collection. The active days and number of tweets features show the different nature of the spam campaigns that promote the domains contained in the two collections.

The spammers indicated by the RBL and the GMF exhibit a longer-lasting behaviour, with an average duration longer by 2.5 days. The top 5%, spammers from these campaigns are active for at least 4 times as many days with a significant increase of 60% for the Active Days metric and a 4-fold increase of the mean value of Active Days. Similar values go for the GMF experiment meaning that RBL spammers rely on massive posts, are more persistent and are active for a longer duration. GMF spammers and legitimate users seem to have the same tweeting patterns (in terms of frequency). *This leads to the conclusion that GMF spammers are most probably regular users with exploited accounts that interject the owner's legitimate tweets with certain spam tweets while RBL spammers have a higher probability of being dedicated spam accounts.*

Some interesting points are also shows by the "Spam URLs" and "Blacklist Hits" metrics that are measures available only for spammers. The Blacklist Hits is the total number

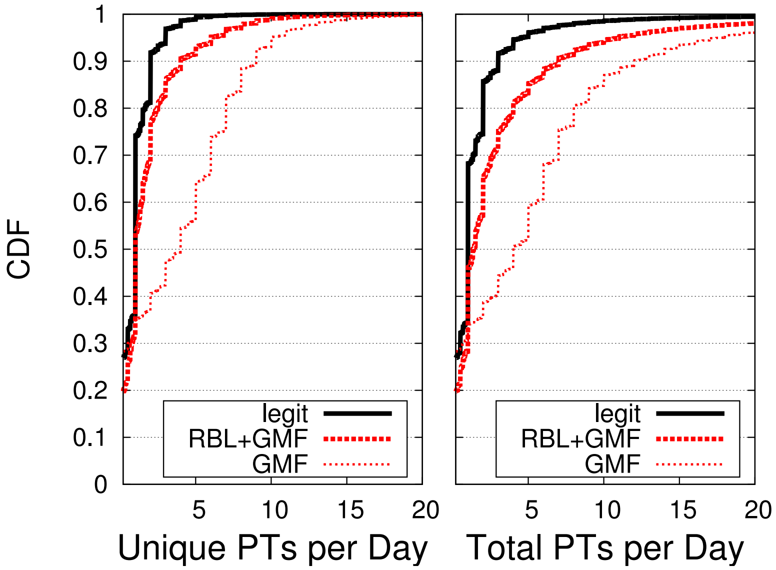


Figure 5.4: CDF of Popular Trends of users weighted by the number of Active Days.

of RBL blacklists that detect a user's spam URLs per day e.g. suppose a user who has posted two spam URLs in a day and the first is flagged by 2 RBLs and the second by only 1, then for this user this metric will be 3. Spam URLs measure is the number of spam URLs posted per day and its average value is similar for both collections (1.6 for RBL+GMF and 1.4 for GMF), which measure is available only for the RBL+GMF collection. On average, a spammer is located by 1.9 RBLs in a single day, according to the results.

After identifying the features that differentiate spammers and legitimate users, we plot the distribution of users for each of those features. As Figure 5.4 shows the distribution of PTs for spammers and legitimate users, we notice on the left subplot the unique PTs per day and on the right one, the total PTs per day.

We notice two different sets of spammers, one for both the RBL and GMF datasets and one for the GMF alone. The distribution of Hashtags in Figure 5.5 are a result of the same procedure. Here we show the extent to which spammers exploit the PTs and Hashtags, in order to promote their campaigns and attract as many users as possible. A key observation here is that unique Trends and unique Hashtags exhibit higher differentiation than the total Trends and total Hashtags.

So we can conclude that the spammers include a higher number of different trending topics and as a result to show up in different user queries. In this way they can achieve a larger coverage and more diverse set of users, while maintaining a constrained approach to the total number of hashtags tweeted in a day. That is an indication of spammers following a stealthy approach and not flooding the system, to avoid being flagged by Twitter's spam detection mechanism. This confirms our initial intuition that posting many differ-

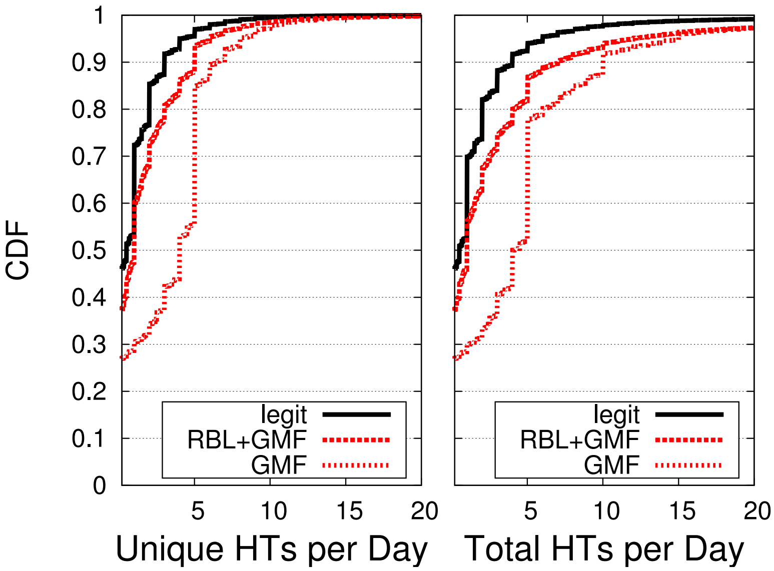


Figure 5.5: CDF of Hashtags of users weighted by the number of Active Days.

ent trending topics is more suspicious than just posting many tweets containing trending topics. Additionally, we notice that GMF spam is far more active in exploiting PTs and Hashtags compared to RBL.

In order to validate if these features (unique Trends and unique Hastags) belong to the same distribution regardless if it is measured in spam or legit users, we performed a statistical analysis (two-sample Kolmogorov-Smirnov test). The p-value (<1e-10) showed that they belong to different distributions.

5.3 The Classifier

In order to classify the spam and legit users we implemented a classification schema described in this section . The class information for each user is the average number of spam tweets per active days. In this way we show not only if a user is a spammer or not, but also the level of spam activity that she exhibits. The features of each user have already been described in a previous section. The only difference is that we normalized the features that are dependent on the number of active days. So all features, except the number of followings and the number of followers. We need these normalizations to build a classifier that is agnostic of the measured period and can be used for an arbitrary time scale.

The following step is to split the dataset into two random subsets, the test and the train set used by the machine learning algorithm in order to assess the predictive ability of the model.

We randomly select the 90% of the initial dataset for the training set and the remaining

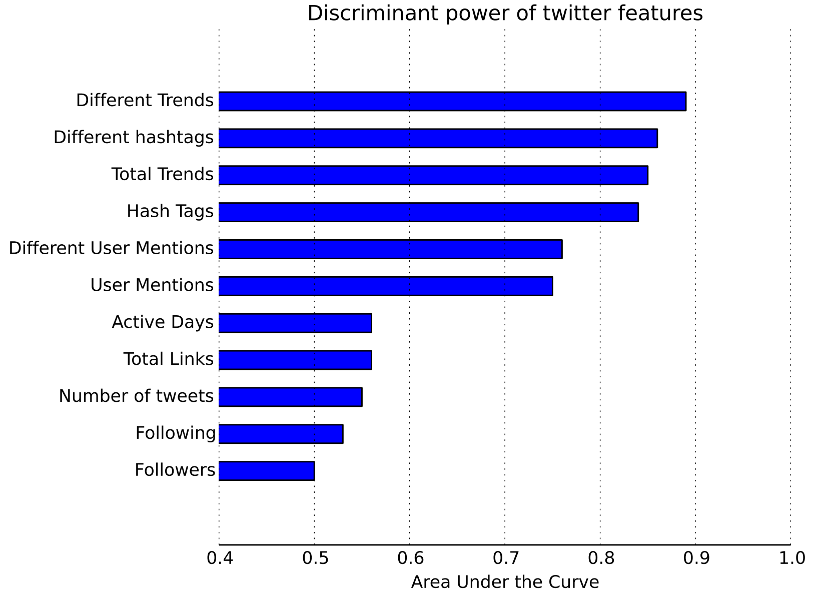


Figure 5.6: The Area Under the Curve (AUC) metric for each feature.

10% as test. Special care was given so that each subset had the same ratio of spam and legit users as the initial dataset, since this random split could create datasets with uneven number of legit and spam users, In [123] they demonstrate that the ratio of spam/legit in a train dataset affects significantly the True Positive Rate (TPR) and the False Positive Rate (FPR). In general, by increasing the spam ratio in the train dataset the FPR is decreased but so does the TPR. Background work has used in the past, in order to tailor the TPR and FPR metrics over or under some acceptable standards respectively, a tactic that we think is unfair. For this reason we kept this ratio the same as in the initial dataset.

The next step after the split of the dataset into training and test set, is to train a Decision Tree Regression (DTR) classifier. Decision Tree Regression is a technique trying to fit a Decision Tree to the train data, by minimising a criterion function. This function in our case is the Mean Squared Error. This is in contrast to traditional decision trees classifiers that trying to fit a decision tree predicting the nominal class of the train data. Consequently, we take into account the rich information that is conveyed in the class feature (average number of spam tweets per active days).

After the training part, we evaluate the TPR and FPR metrics of the classification on the test dataset. We apply a repeated random sub-sampling validation as follows: we repeat this procedure 100 times and for each time we build a novel train and test dataset as presented above, we train a DTR and assess the TPR and FPR metrics. The final step is to report the average TPR and FPR metrics. For this we used the python package scikit-learn [249] to train and assess the DTR classifier.

We apply this classification schema on the tweets dataset, collected in March 2014 containing 622,428 users, from which 5,152 have sent at list one GMF spam tweet. These users have posted 6,658,282 tweets, totally. The average TPR was 75.2% (95% CI [74.8, 75.5]) and the average FPR was 0.25% (95% CI [0.246, 0.255]). Finally, we analyze the discriminatory ability of each feature, by measuring the performance of the classification scheme when we apply only one of each feature and we blind the rest. The next step was to measure the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) plot of the trained model. When AUC is equal to 1.0 indicates a perfect classifier, while 0.5 indicates that the model does not perform better than a random classifier. This procedure revealed that both total, different trends and hashtags perform good as features for spam detection (Figure 5.6).

5.3.1 Tweets classification

As shown above, the presented features exhibit a fairly good ability to classify spam users with a TP rate of 75.1% and FP rate of 0.26. This section presents the ability of the features to discriminate individual tweets that belong to the GMF campaigns. In this procedure, a negative class was assigned to every tweet that does not belong to the GMF campaign. We use the same dataset as before (dataset of March 2014) containing 63,612 tweets than belong to the GMF campaigns (out of 6.6 million). We extracted the following features, for each tweet: Total Links, Hash Tags, User Mentions, Retweet Status, and Total Trends. We use the same learning and validation schema as used in the user classification procedure. The only difference is that we applied a Decision Tree Classifier (DTC) instead of a DTR because of in this case we are facing a binary classification problem and DTC is more applicable to this task.

The classifier achieved an average TP rate of 81% (95% CI [78.2, 82.7]) and an average FP rate of 0.58% (95% CI [0.56, 0.6]). In comparison, [123] succeeded 55% TPR and 0.4% FPR on Twitter data and 80.8% TPR and 0.32% on Facebook data. Of course, identifying spam that belonged in the GMF campaign is a relatively easy task since it is based largely on hijacking of trending topics. Since there are other spam campaigns (beside GMF) that are exploiting trending topics, it is fair to assume that a significant proportion of the False Positives are actual Positives that belong to similar spam campaigns.

The rule here is that if Total_Trends <3.5 then it is legit else it is spam, since the root node is the most discriminant rule in decision trees. This single rule has 79.6% TPR and 0.77% FPR, which practically means that just by checking the number of Trends in a single tweet, an algorithm can identify 4 out of 5 tweets that belong in the GMF campaign. The rest 1 out of 5 are False Negatives.

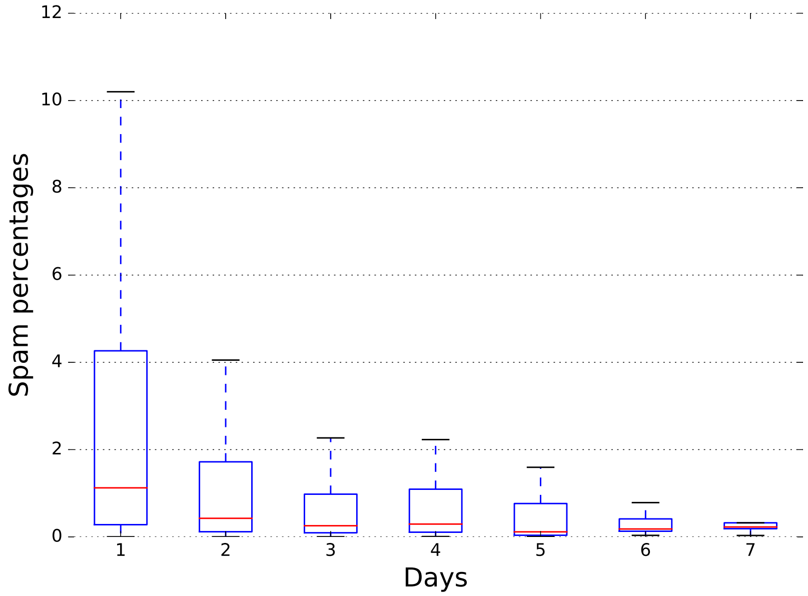


Figure 5.7: Boxplot of spam percentages.

5.3.2 Time scale analysis of Popular Trends

Additionally, we analyze the time that it takes for a Hashtag to be exploited, after it becomes a PT. We use the collected dataset of March of 2014 containing 5,780 different PTs. *The part that had been used in at least one spam tweet was 3,193 (55%).* The 74.5% of these trends where involved in spam tweets for only a single day and 15% for two days. *Consequently, we can say that the vast majority of PTs are exploited the very first days of their creation. It is very likely that spammers add to their spam tweets PTs that are acquired from Twitter daily.* In figure 5.7 we show the boxplots of spam percentages in tweets containing PTs that lasted from 1 to 7 days. From this analysis does not include the trends that lasted more than 7 days since they were very few (<1%). *It is impressive that 1% of tweets that contain trends that are active only for a single day is spam.*

We also notice a downward trend, which means that PTs that are active for more days contain a lower percentage of spam. Nevertheless, we observe an upward trend (figure 5.8), if we repeat the same plot with the absolute number of spam (rather than percentages). This indicates that PTs that are active for more days contain a higher amount of spam. As we have seen PTs that are active for more days are more generic, thus they are exploited in a higher level. Nevertheless, targeting PTs with shorter duration is more effective since it is exploiting novel trends and potential unsuspicious users.

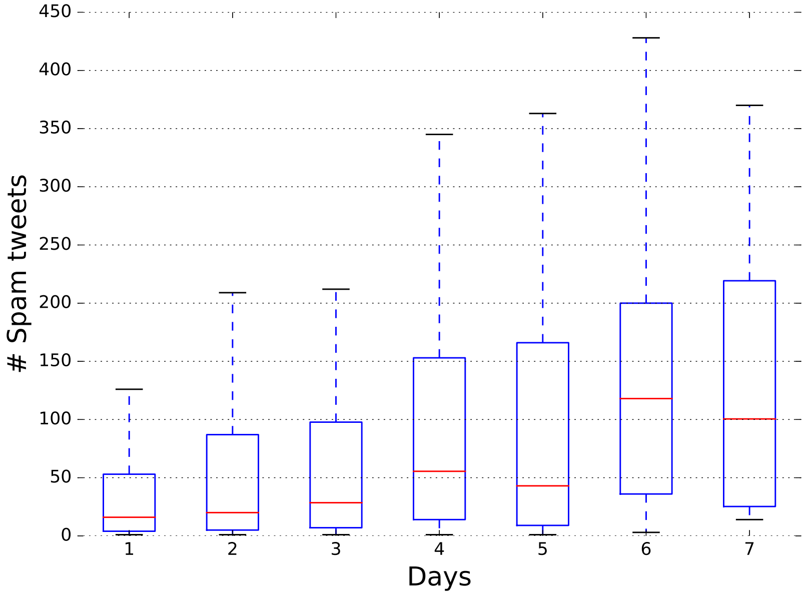


Figure 5.8: Boxplot of Number of Spam tweets.

5.4 Spam campaigns

As we have previously discussed, in most cases spam delivery happens on a massive scale with well orchestrated behaviour [301]. To further illustrate this, we plot a recreation of a spam campaign that was captured on a single day (January 10th 2014) in Figure 5.9. This campaign comprised of 17 different spam domains belonging to the GMF class, which involved 1,604 different users. 16 out of these 17 domains belong to a single IP address. This IP also hosted 2 more domains that were contained in the blacklists, but that did not take part in this campaign. The average edge degree for a node that represents a spam domain is 125, i.e., every spam domain in the graph could be found in average, in the tweets of 125 different users. In contrast, the average edge degree for the graph depicting users and the URLs they tweeted for that day is 2.3 for legitimate URLs. Various graph properties can be exported either from spam campaign graphs, or user-URL graphs, that can potentially assist in identifying spam users and campaigns. We plan to explore this as part of our future work.

The most successful domain is tweeted by 337 users, while the least by 40. These numbers demonstrate the stealthy approach of this type of campaigns that do not flood Twitter with multiple messages advertising the campaign, which would result on being detected by Twitter's spam detection mechanism. The average number of tweets per user ranges from 1.07 to 1.58, meaning that in the worst case only half the users will promote the spam domain a second time.

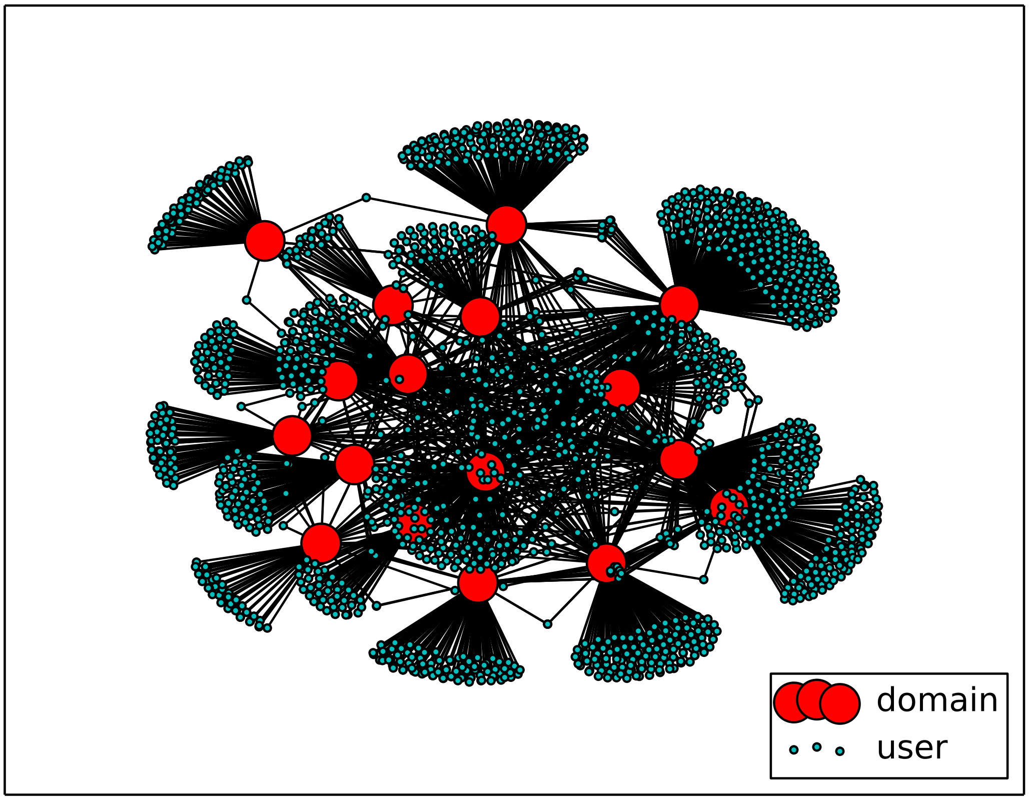


Figure 5.9: Plot of a spam campaign involving 17 GMF domains and 1,604 users

5.4.1 URL based campaign detection

The form of a posted link (or URL) can give valuable information regarding the possibility to target spam sites. For example masqueraded links or links that have been shorted from multiple URL shortener services.

Another valuable source is the graph that shows the users that posted these links. To demonstrate this we collected tweets with popular trends for a single day (10 Jan 2014). In total we collected tweets from 460,000 users that have 110,000 links. Consequently, we collected all links and we discarded links that have been posted by more than 2 users. On Figure 5.10 we plot users as nodes. A user is connected to another user if they share a common URL. This plot shows the high connectivity of twitter users based on the links that they post. We observe that there is a large cluster of users at the centre of the graph. That means that a high proportion of users are connected with links, even if these links are not famous at all (have been posted by 2 users at most). We also notice structures (like super-users or users that are connected with disproportionately many other users) that we see usually in follower/followings plots. If we repeat the same plot but we allow higher connectivity (allow links that have been posted by more than 2 users) we will notice an exponential growth on user connections in a degree that visualization becomes very difficult. Nevertheless this figure is of great importance because it demonstrates that users can be heavily connected even when looking for factors that seemingly should not connect many users (like links that have been posted by no more than 2 users).

In figure 5.11 each node is a URL. The size of the node is proportional to the number of users that have posted this URL. Two URLs A,B are connected if the group of users that posted A and the group of users that posted B are common in a factor of 50%. For example if URL A has been posted by 200 users (group A) and URL B has been posted by 70 users (group B) then we notice that they have 50 users in common (50 >70 * 50%). This is how we define that a two user groups share a common interest. In other words, figure 5.11 shows the URLs that the groups that posted them are related. In normal (not spam URLs) we expect that we would see a very few sets of URLs connected given these strict criteria. Nevertheless we notice that there is a cluster of URLs in the center of the graph that all of them share common interests. That means that these URLs are semantically related. Indeed a closer inspection of these URLs proved to belong to the Get More Followers campaigns. This phenomenon is known as link farming [128]. According to this, spam accounts that belong to a specific scheme follow each other, in order to achieve high number of followers and consequently appear as more trustworthy and gain additional influence. This phenomenon is also common in the web as a method to increase ranking in search engines. Formally, link farms are cliques in the social network, although any set of nodes with unusual high density of links between them can be an indication of link farming. This figure demonstrates how we can identify abnormal, or suspicious substructures in social graphs that indicate malicious behaviour. Of course these structures do not have to be always malicious, but they are worthy to be researched in order to investigate the behaviour of sub-communities in Social Networks [165].

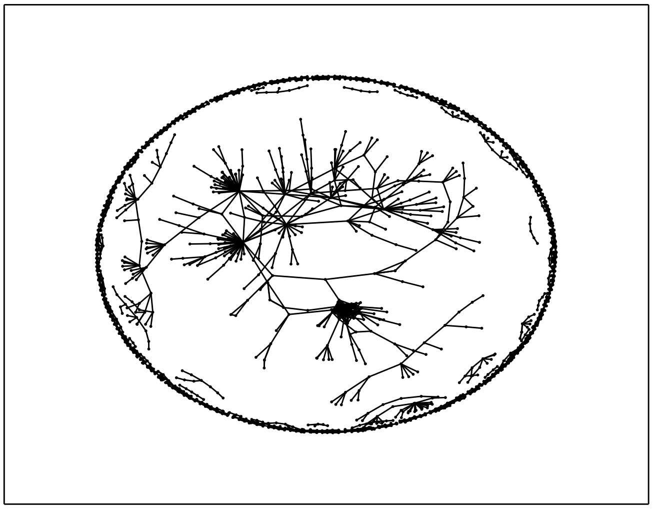


Figure 5.10: Users connected if they have posted same URLs.

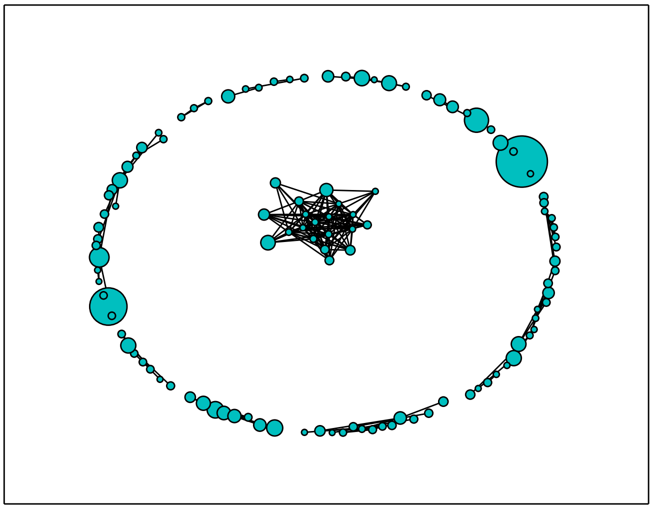


Figure 5.11: URLs are connected if the users that posted them are at least in a level of 50% common. The network at the middle are links belonging to the GMF campaign.

Chapter 6
Study through NLP and Sentiment analysis in Twitter

In this chapter we demonstrate a study of topic analysis study of a dataset with political content in Twitter, by applying NLP techniques. The main scope is recognize tweeting patterns, sentiments and the semantic relations of the most important entities that prevailed during the online discourse that preceded these two events. To accomplish this, we collected all tweets referring to these two events and we applied Named Entity Recognition and performed a comparative analysis of the number and sentiment of tweets regarding different political parties, politicians and institutions. Next we show the temporal sentiment variation for the major parties and politicians. We compile a novel sentiment dictionary for the Greek language and that we also account for the presence of sarcasm that has been found to severely confound sentiment analysis [205]. We also use Latent Dirichlet Allocation, which is an unsupervised learning method that estimates the probability of an entity to belong to a distinct cluster, also called "topic". We were able to assess the semantic proximities of political parties, politicians and major institutions. The results of this analysis revealed part of the public sentiment towards main entities along with their semantic proximities. Also we show that there was a strong anti-austerity sentiment accompanied with a critical view on European and Greek political actions.

6.1 Methodology

This chapter describes a natural language analysis on a Twitter corpus related to two electoral events on 2015, during a politically turbulent period of Greece that was triggered by an effort to negotiate a reconstruction of its national debt. These events were the Greek bailout referendum that took place at 5 July 2015 and the second was the subsequent legislative elections that took place at 20 September of the same year. The aim of this study is to identify the tweeting patterns, the expressed sentiment and the semantic relations of the most important entities that prevailed during the online discourse that preceded these two events.

Initially we split our analysis in 5 distinct parts. The first part consists of the data collection and Entity Identification (EI)s. EI is defined as the process of extracting the most important notions (entities) that are prevalent in users' posts, i.e. "Prime Minister" and "Debt". Next each entity is represented by a set of words with equal meaning (i.e "EU" and "European Union"). Although there is a variety of methods for automatic EI, they are all inferior to various extend to human curators [94]. For this reason and given the relative narrow semantic context (elections) of our dataset this task was performed manually. Having in our disposal a set of prevalent entities we proceeded to perform Volume Analysis. The Volume Analysis studies the count differences between tweets that belong to certain entities.

Contradicting studies about the predictive ability of tweets count for election results (for examples of positive findings see [280,309] and for a negative see [146]). Nevertheless, most of these studies conclude that tweet count can give valuable information if not for the election outcome then for the quantitative estimation of the political inclinations of Twitter's user-base. Given also that the referendum dataset has simple structure (YES/NO), we apply Volume Analysis.

The results show that indeed tweet counts matched the referendum results and we also associate changes in the temporal variation of the ratio between "YES" and "NO" tweet counts with real events. The next part of the analysis include the study of Entity co -occurrence. Here we visualize in 2-D space entities by simulating a graph of spring forces. The graph shows that the higher the number of co-occurrence between two entities the stronger the force. In [228] they use this technique to visualize (among other) semantic data and online social networks [258] but it has not been used, to our knowledge, to visualize Twitter extracted entities. The advantage of the technique is that when when applied to the referendum dataset, it is a very computationally efficient method that gave insights on the "NO" and "YES" affiliated entities . The fourth part of the analysis is the sentiment analysis and perhaps the most notable collection of methods for analysis of textual content that is rich of human opinions [244, 246]. The novelty of our study is that we use a novel sentiment dictionary for the Greek language and that we also account for the presence of sarcasm that has been found to severely confound sentiment analysis [206]. The temporal variation of sentiment for various entities along with the identification of the most and least sarcasm-prevalent entities provided additional insights on user's opinions.

The last step includes the topic modeling, an unsupervised learning method estimating the probability of an entity to belong to a distinct cluster, also called "topic". Each topic is an automatically-extracted semantic structure of the input corpus. Topic modeling can help us extract the hidden semantic similarities of our data and visualize their proximities in 2-D space.

Topic modeling algorithms like Latent Dirichlet Allocation in Twitter data entails certain difficulties due to the brevity of text messages [192] that we were able to overcome by applying sarcasm correction. Topic modeling revealed that "YES" and "NO" entities were unexpectedly close in the referendum dataset. It also spatially outlined the relationships of political parties that took part in elections.

Additionally ,the application of these methods to Greek tweets entails some additional difficulties. The online language, primarily used by the youth, is a mix of Greek grammar with Latin letters, called "greeklish" and stems from early text-based communication systems that had limited support for Greek letters. This system often disregards Greek grammar and punctuation and has not any standard correspondence between Greek and Latin letters, resulting in a highly complex language with multiple possible writings even for basic and short words that makes automatic detection a very tedious task [76], [219]. Although nowadays the majority of Greek users are using Greek characters when tweeting, most of the included hashtags are present in a "greeklish" form. Also the demographic subset of Greek Twitter users is narrower than in other western countries thus limiting its representative power [72]. In [72] they perform irony detection in Greek political tweets and inferred similar percentages with studies focusing in U.S. politics, despite of these difficulties.

Political Background in Greece

The new "anti-austerity" government of the SYRIZA party was elected on 25th of January 2015 in Greece with a percentage of 36.3%, starting a long negotiation with the Eurogroup about debt reconstruction. There was no visible progress achieved until June 2015 and the government of SYRIZA decided to throw a referendum on 5th of July 2015. This way the Greek people would decide whether to accept or not the current austerity measures proposed by the Eurogroup. The result was that capital controls were enforced in Greece and the result of the referendum was NO (do not accept) with a percentage of 61.3%. The result of the referendum was not accepted by Eurogroup as a bargaining tool and under extreme pressure, the government decided to accept the proposed measures. Several disagreeing members of the SYRIZA party threatened to vote down the measures. The prime minister (Alexis Tsipras) decided to expel the disagreeing Members of the Parliament that belonged to the governing party, and announced new legislative elections on 20th of September 2016. The next elections gave the victory to SYRIZA again with a reduced percentage of 35.5% and the party formed mainly from disagreeing members (called LAE) did not get enough votes to enter the parliament.

6.1.1 The dataset

The dataset used in this study is consisted of two parts: the first one are all tweets containing the #dimopsifisma and #greferendum hashtags. The most prominent hashtags that

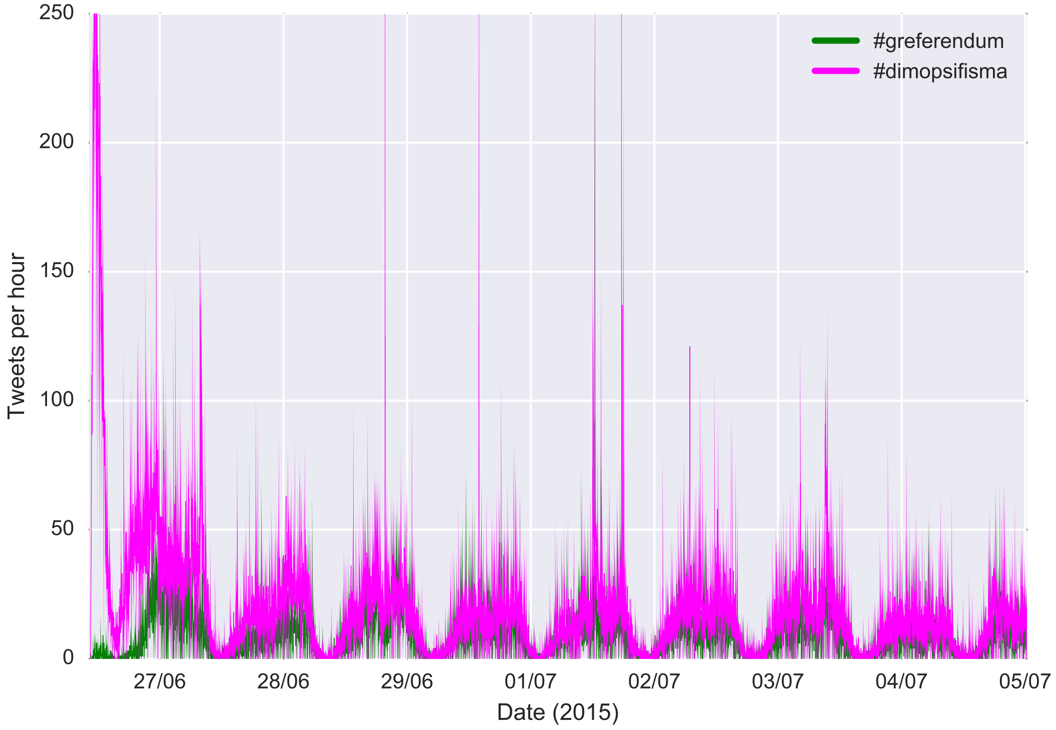


Figure 6.1: Frequency of referendum tweets per hour. The frequency of tweets peaked right after the Referendum was announced and followed a small declining trend. The day/night patterns are also visible.

prevailed throughout the period that preceded the Greek bailout referendum ("dimopsifisma" is the Greek word for referendum), were by far these. The data were collected from 25th June 2015 when the referendum was announced, until 5th July 2015 when the referendum took place, through Twitter's API. Totally 301,000 tweets were contained in this dataset, out of which 84,481 are neither retweets nor replies. In figure 6.1, we show the frequency of referendum tweets is, where we notice the day and night patterns as well as a decline of tweets frequency over time.

The second dataset contains all the tweets containing the hashtags #ekloges and #ekloges_round2 ("ekloges" is the Greek word for elections), which dominated the online discussion regarding the Greek legislative elections that were announced on 20th August 2015 and were held on 20th September of the same year. This dataset contains 182,000 tweets, totally, out of which 45,750 are neither retweets nor replies. These two datasets contain mainly the complete online discourse that happened in the Greek Twittersphere regard-

Table 6.1: Entity Variants from Plain Text and Hashtags

	Text Entities	Hashtag Entities
Number of Unique Entities	156	116
Min Number of Variants per Entity	1	1
Max Number of Variants per Entity	74	148
Average Number of Variants per Entity	18.9	21.6

ing the two political events (referendum and elections). The analysis considers only tweets with at least one Greek letter. This filtering is essential to eliminate content not representing the Greek electorate, since these events had attracted a worldwide interest (especially the referendum).

6.1.2 Entity Identification

We performed entity identification [79] on the elections and referendum Twitter corpus, in order to support our analysis and reveal relationships between persons, institutions, events and abstract notions (such as democracy or liberalism), As a first step, we gathered all unique words and Twitter hashtags present in the tweets along with the respective occurrence frequency of each. Next we selected manually all entities relevant to the political domain of the Greek legislative elections and referendum of 2015, apparently considering the most frequent words and hashtags as of higher importance. We grouped all various forms that a given entity appears in, so that we would be able to identify a certain entity regardless of the variant it appears with in the tweets. For example, all of the following hashtags identify a single entity, that of the Greek Prime Minister, Alexis Tsipras: #Tsipras #atsipras #alexistsipras #atsipra #aleksitsipra. Finally we group the variants for entities found either as plain text in the tweets or mentioned as hashtags.

We extracted 156 entities totally, from plain text of tweets and 116 entities from Twitter hashtags; the minimum, maximum and average number of variants per entity is listed in Table 6.1. We performed a normalization of all tweets in order to minimize variation coming from common spelling mistakes, before matching entities appearing in hashtags and in the tweet text. So we grouped commonly misspelled diphthongs and punctuation into a single form. Subsequently, in order to improve precision of entity identification, we linked and combined occurrences of a given entity that appeared both as plain text and as hashtag. Finally, we located all entities that are referenced either as hashtag or as plain text for each tweet, distinctively, and annotated our dataset accordingly for further processing. In order to import semantic knowledge around the context of these events into our analysis, this laborious manual task was necessary.

6.1.3 Sentiment Analysis

We used SentiStrength [10], that is ideally suited for the affective dimension of the social web and Twitter in particular [298] in order to perform sentiment analysis. Texts often consists of a mixture of positive and negative sentiment and for some applications it is necessary to detect both simultaneously and also to detect the strength of sentiment expressed. SentiStrength extracts positive and negative sentiment strength from short informal electronic text by using several methods. The main power of SentiStrength is in the combined effect of its rules to adapt to various informal text variations as well as in the overall approach of using a list of term strengths and identifying the strongest positive and negative terms in any comment. It introduces a dual 5-point system for positive and negative sentiment and reports two sentiment strengths associated with a given piece of text: -1 (not negative) to -5 (extremely negative), and 1 (not positive) to 5 (extremely positive). Because humans process positive and negative sentiment in parallel, SentiStrength uses two scores. It is reasonable to conceive sentiment as separately measurable positive and negative components, since positive and negative sentiment can coexist within texts, e.g. the text "I love you but hate the current political climate." has positive strength 3 and negative strength -4.

6.1.4 New Lexicon for Sentiment Strength Detection

It is known that sentiment analysis domain-dependent which means that applying a classifier to a dataset different from the one on which it was trained often gives poor results [35]. Indeed, the diversity of topics and communication styles in the social web suggests that many different classifiers may be needed. The political domain may require individual treatment while the existing general-purpose social web sentiment analysis algorithms may not be optimal for such texts focused around specific topics. SentiStrength's disadvantage is that its general sentiment lexicon performs poorly and achieves very low accuracy in political texts. SentiStrength however supports topic-specific lexicon extension involving adding topic-specific words to the default general sentiment lexicon [297].

For these reasons we created a new general-purpose and political-domain lexicons through manually selecting and annotating words from the Twitter corpora, to enriched SentiStrength for the Greek political domain. Human intervention seems likely to be particularly important for narrowly-focused topics for which small misclassifications may result in significant discrepancies if they are for terms that are frequently used with regard to a key aspect of the topic. We manually created a new SentiStrength-compatible lexicon comprising Greek words with associated positive/negative sentiment strength, aiming to improve the accuracy and effectiveness of political-domain lexical sentiment strength detection. The SentiStrength algorithm uses direct indications of sentiment for sentiment strength detection across the social web primarily. We included indirect affective words too, in order to enhance sentiment detection, since our study is domain-dependent and time-dependent (political domain and Greek legislative elections and referendum of 2015, respectively). These words do not directly express the terms, but rather identify them and associate with sentiment.

The new sentiment detection lexicon we compiled, is a merge of the following 3 lexicons in Greek: (i) SentiStrenth's built-in lexicon that provides general sentiment analysis for the Social Web; (ii) SocialSensor lexicon, utilized by SocialSensor framework to collect, process, and aggregate big streams of social media data and multimedia to discover trends, events, influencers, and interesting media content in real time [307]; and finally (iii) our new political-domain lexicon introducing lexical sentiment strength detection for political texts, and is based on frequently used terms in the elections and referendum Twitter corpus. There may be rare words or specialist words that are frequently used to express sentiment, for a given topic. We identify these words through our lexical extension method and use them to improve sentiment strength prediction through a political-domain lexicon extension (i.e., a set of words and word strengths). In Table 6.2 we see the size of lexicons expressed as number of words contained. Last but not least, SentiStrength is covering the word inflections, since it allows for insertion of wildcards (an asterisk character *) at word stems in the lexicon. In this way we have made extensive use of wildcards in the newly created sentiment strength detection lexicon. This feature is extremely useful to enhance word matching and is particularly suited to Greek language morphology, since Greek is a highly inflected language.

Table 6.2: Number of Words Contained in Lexicons

	Number of Words
SentiStrength	1638
SocialSensor	2315
Domain-Specific	974
New Lexicon	4915

6.1.5 Sarcasm Detection

As mentioned before, there is a significant part of (~50%) of tweets referring to political issues that is sarcastic or humorous nature and can severely obscure the analysis, particularly in the political domain [92], [262]. In order to identify this content, we apply a method used by the online sarcasm detection service, [3], in order to be able to characterize Greek text.

Initially we needed to construct a sarcasm classification mechanism [137], [189], so we built a database containing all the original tweets. These were the tweets that were neither retweets nor replies. From 483,000 tweets totally belonging to the referendum and election datasets, we extracted 130,231 original tweets. Next we built a website that showed random tweets and users got to choose whether each tweet was sarcastic or non-sarcastic/normal. Users could also skip a tweet in the case that they could not make a safe decision. The website also explained the context of the study including a simple explanation of "sarcasm" in Twitter, in order to assure uniform classification from many human judges.

We promoted the website through social media and after a week we collected the human-flagged tweets. In cases of conflicting flaggings, we applied a unanimity rule. Namely we removed from the ground truth all tweets with conflicting flaggings. This resulted in 2,642 tweets flagged as sarcastic (positive) and 2,002 were non-sarcastic/normal (negatives) tweets from 134 different user sessions.

Using this human-flagged dataset regarding sarcasm, we continued to build a classification model. We extracted lexical and semantic features, from 4,644 flagged tweets, totally. We developed a stemmer for the Greek language,in order to build the lexical features and built a stopword collection containing commonly used Greek words. Next we extracted 1-grams and 2-grams for each tweet.

We do not use Part-Of-Speech (POS) since we could not locate an adequate POS dictionary for the Greek language. The semantic features that we included were average sentiments for each word in the tweet and topics. Also we used the same SentiStrength-compatible dictionary for sentiment that we constructed for the purposes of this study. We generated 100 topics related to the context of the collected tweets by performing topic analysis. The hypothesis here is that some topics are more associated with sarcastic tweets and therefore, by incorporating them as features, we can improve the classification efficiency of our model. In order to perform topic analysis we used Latent Dirichlet Allocation (LDA), implemented with the Gensim Python library. We used a Support Vector Machine (SVM) classifier with a linear kernel and an Euclidean regularization coefficient of 0.1, for classification. Next we randomly divided the flagged dataset into 70% training dataset and 30% test dataset, trained our model and estimated its performance on the test dataset. The classification results are on Table 6.3.

Table 6.3: Classification Results

	Precision	Recall	f1-score	Test Samples
Non-Sarcastic	0.69	0.62	0.65	621
Sarcastic	0.72	0.78	0.75	772
Average/total	0.70	0.71	0.70	1393

Also Charalampakis [72] performed sarcasm detection in Greek tweets regarding politics and reported 80% True Positive Ratio (TPR), but with extremely low number of samples (126). Our technique results in a TPR estimate of 0.78 for sarcastic tweets is similar to estimates from other studies, such as 0.71 by Gonzalez-Ibanez [137] and 0.75 by Liebrecht [189]. In the political context, "sarcasm" is a subtle and ambiguous notion so it is questionable whether significantly superior results are possible. This conclusion is supported by the fact that even humans have a limited ability to detect sarcasm in Twitter that ranges from 70% [137] to 85% [189].

We generated "sarcasm values" for all 130,000 original texts in our dataset, using the trained SVM classifier. The values have the form of percentages ranging from -100% (definitely not sarcastic) till 100% (definitely sarcastic). The SVM classifier calculates a confidence score, for each tweet, which is the signed distance of that tweet from the optimal hyperplane calculated during training. Then we applied the hyperbolic tangent ($\tanh(x)$) as a sigmoid function to convert this distance to percentages. Finally, we mapped each "sarcasm value" to one of the following categories: "no_sarcasm" for negative values, "sarcasm_1" for values from 0% to 20% of positive sarcasm values, "sarcasm_2" for values from 20% to 40% of positive sarcasm values, and "sarcasm_3" for values greater than 40% of the "sarcasm value".

The sarcasm detection revealed interesting indirect affective words, which are words used mainly in sarcastic tweets for mocking or ironic purposes. The top indirect affective words were ATM (due to Capital Controls, people could withdraw a limited amount of cash through ATMs), Hope (used in SYRIZA's slogans), Merkel (Germany's Chancellor), memorandums (sets of austerity measures imposed by EU), bankruptcy, drachma (Greece's currency before Euro) and recovery.

6.2 Results

6.2.1 Tweets' Volume Analysis

The tweets' volume can give insights regarding specific events, although it is not a sufficient indicator of political inclinations of users, it can give insights regarding specific events. Figure 6.2 shows the volume of referendum tweets per hour. We only focus on tweets containing either *voteYES* or *voteNO* entities. The spikes in this plot are indicative of major events during the pre-referendum period. We revealed though text analysis that these tweets were either prompting people to participate in certain demonstrations or they were retweets of the prime minister, urging for "NO" votes. In figure 6.2 we also show the decreasing temporal variation of the ratio of users who included "YES" vs. "NO" entities in their tweets. The number of "NO" tweets were persistently higher than "YES" tweets throughout the pre-referendum period. Also certain "NO" promoting tweets and events generated a public sensation that are visible as spikes in the red line.

Surprising, the opinion polls conducted during the same period showed an opposite

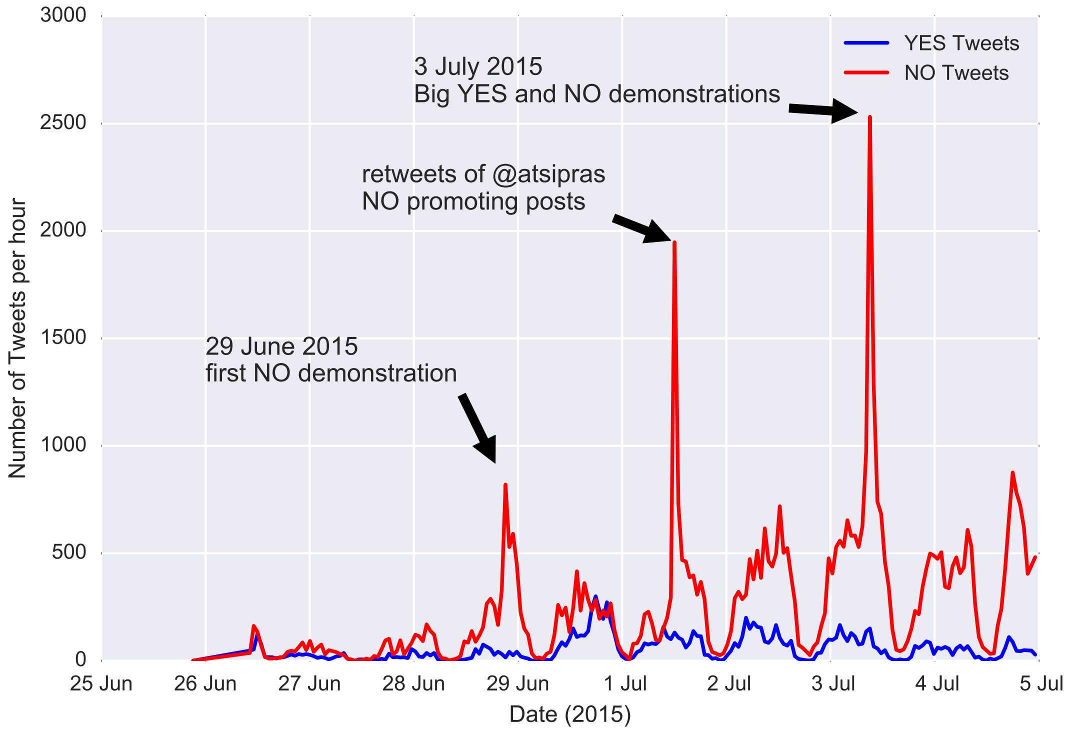


Figure 6.2: Frequency of YES/NO tweets in Referendum.

trend, which, according to post-referendum analysis, was erratic [5]. Despite the high difference from the final result (38.6%), the final "YES" vs. "NO" ratio right before the referendum was 18%, which, was very close to the preferences of the demographics of Greek Twitter users. Users belonging to the age groups of 18-24 and 25-34 voted "YES" with a percentage of 15% and 27.7%, respectively [2]. Figure 6.3 shows the effect of Capital Controls on the "YES" vs. "NO" ratio. We assume that the enforcement of Capital Controls temporarily strengthened the "NO" sentiment. In this plot red and blue lines represent the cumulative number of users that have posted exclusively "YES" and "NO" tweets respectively for each time point of the pre-Referendum period. The black dashed line is the "YES" to "NO" user ratio and the solid black line is the final "YES" percentage (38.6%).

As mentioned above, the dataset for the referendum contains all the tweets from the most prominent hashtags (#dimopsifisma and #greferendum) from 25th of June 2015 when the referendum was announced, until 5th of July 2015 when the referendum took place. This dataset contains 301,000 tweets, out of which 84,481 are neither retweets nor replies. In figure 6.3 the users are selected according to the entity they belong, during the entity extraction phase, described in chapter 6.1.2. This means that we select only the users that belong to the VoteYES and VoteNO entities. We do not include users that have included both voteYES and voteNo in theirs tweets (they belong to VoteYES and VoteNO entities) only users belonging to one of these two entities. The total number of users for the referendum was 40748 and from these we selected only the users that belong ONLY to one of the VoteYES or VoteNO entities, 11.672 NO users and 1558 YES users, making in total 13230 users for both entities.

Traditional election polls in Greece, survey, at best, approximately 2.000 people. For example in one of the largest opinion polls regarding the referendum [23], they surveyed 2,000 people. Therefore, we state that Twitter polls have a larger number of samples (samples = 13,230) compared to traditional polls (sample ~= 2.000). Of course Twitter polls suffer from selection bias since these users do not consist the whole electorate for Greece, for this event. It only reflects the representation of the voters that post in Twitter, who according to many studies belong to a very specific user group, age group, ideologically and social. But for this group we have the result, which states that it is was 15% - 27.7%, (our result was 18%). For the dataset we collect only the tweets with at least one Greek letter. In this way we do not take into account the tweets for the international users that tweeted, but do not belong to the electorate.

The volume of tweets referring to the leading party (SYRIZA) and its leader (Alexis Tsipras) had a decreasing trend during the pre-elections period (figure 6.4). In opposition, we notice a slight increase in the volume referring to the SYRIZA's major opposition party, New Democracy (ND). Nevertheless this volume of leading party (SYRIZA) remained higher than the tweets referring to the main opposition (ND) and its leader (Meimarakis). Also the total number of the pre-elections tweets (180,000) that lasted for one month, was significantly lower than the pre-referendum tweets (308,000) that lasted for only one week. This is supported by the fact that the elections turn-out was exceptionally low (56.6%). Consequently, the tweet volume forms a good indicator of the general enthusiasm or apathy feeling towards the elections, although there were not very strong variations in sentiment. Also, the tweet volume can give a precise estimation of the final result for the demographics that Twitter represents, when the predicament of a referendum is simple (like a "YES"/"NO" question).

An additional interesting question regarding tweets' volume is the potential existence of different tweeting pattern between "YES" and "NO" voters. We measured the average tweets posted by "YES" users and "NO" users in order to check this. "YES", is defined as a user who has posted at least one "YES" entity and none "NO" entity. Similarly, "NO" users are defined accordingly. Our dataset had 1.558 "YES" users in total and 11.672 "NO" users. Still, in average, "YES" users sent approximately twice as many tweets (11.3) than the "NO" users (6.1).

Overall we conclude that, "YES" voters posted more tweets than "NO" voters and used more their respective hashtag (#VOTEYES) than "NO" voters used their respective hashtag

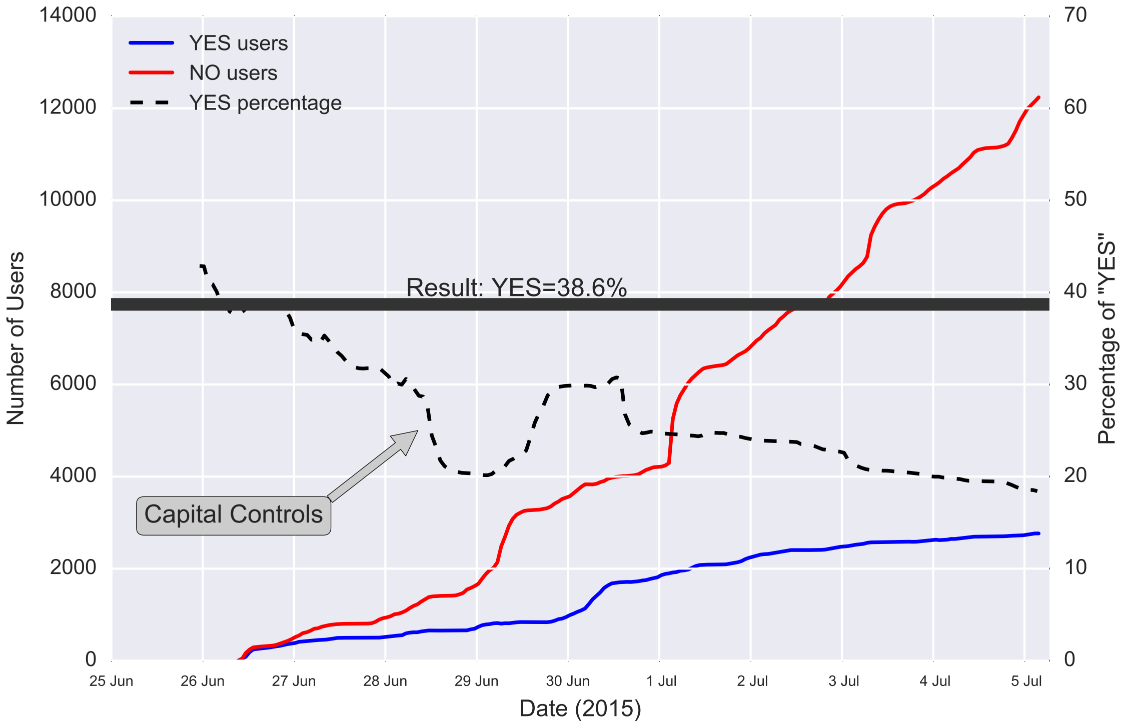


Figure 6.3: Variation of YES percentage.

("#VOTENO"). We support this statement with a statistical analysis which rejects the hypothesis that this observation is a random finding. Or else, certainly, there was an effect that guided this difference in the averages. There could only be two types of effects to justify this finding. The first is that "YES" users were engaged in an orchestrated campaign to promote "YES" content. This of course does not exclude the possibility that "NO" users did not take part in similar campaigns. However, if we set as campaign objective the posting of many tweets containing a set of party related hashtags, then the "YES" campaign was more effective. The second is that "YES" users exhibited a normal tweeting behaviour, whereas "NO" users showed higher apathy both at tweeting activity and hashtag use. Since we do not have a baseline of which is the "norm" in party-related hashtag use from a given group of supporters during this electoral event, we cannot make safe deduction regarding the nature of the effect that took place. It is important also to note that by "#VOTEYES" and "#VOTENO" we refer to a set of hashtags promoting the "yes" and "no" vote respectively. Overall we applied two statistical tests with two null hypotheses:

- The number of tweets from YES voters comes from the same distribution as the num-

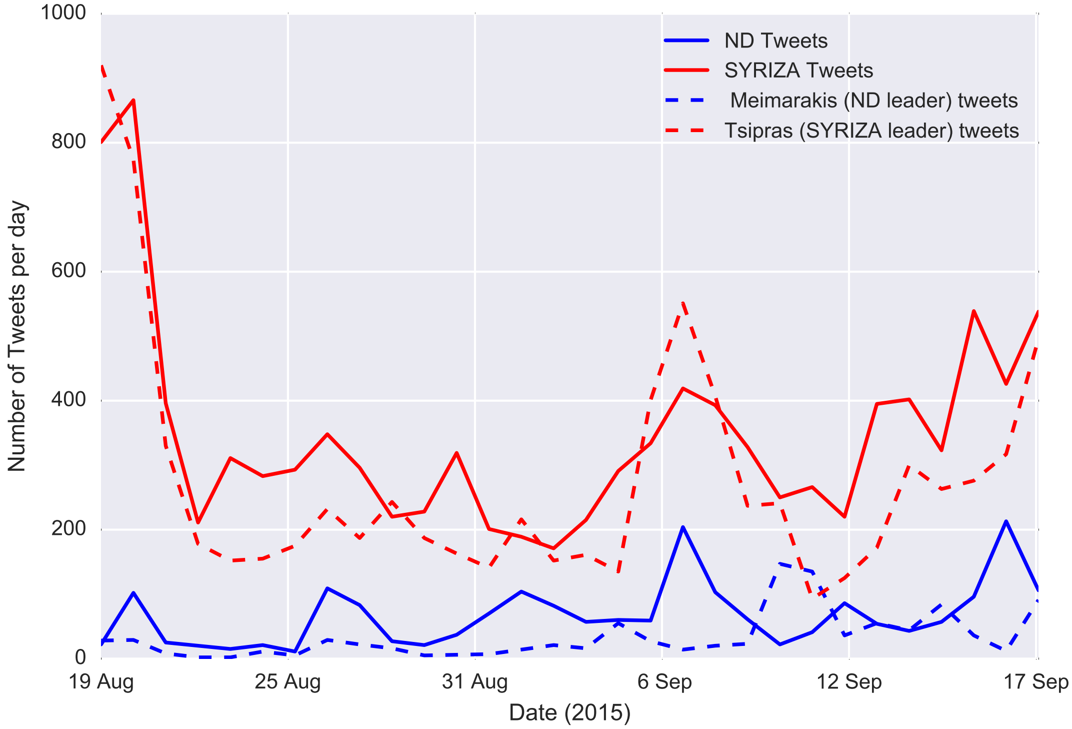


Figure 6.4: Frequency of Election tweets.

ber of tweets from NO voters.

- The number of tweets containing the "#VOTEYES" hashtag from YES voters comes from the same distribution as the number of #VOTENO tweets in NO voters.

Another way to formulate these two null hypotheses is: "If we take a random YES voter and we count the number of tweets that she has posted (say A) and a random NO voter and we count the number of tweets that she has posted (say B), then the probability of A being greater or equal than B is equal with the probability of A being less or equal than B (or else: P(A<=B) = P(A>=B))". Similarly: "If we take a random YES voter and we count the number of tweets containing the "#VOTEYES" hashtag that she has posted (say A), and a random NO voter and we count the number of tweets containing the "#VOTENO" hashtag that she has posted (say B), then the probability of A being greater or equal than B is equal with the probability of A being less or equal than B (or else: P(A<=B) = P(A>=B))". We applied the Mann–Whitney U test which is a non-parametric test. This means that we don't have to make any assumption regarding the underlying distribution of the data. The only prerequisite of the test is that the two measurements are independent which is obvious since the two samples (YES voters and NO voters) are completely disjoint. In both cases the p-values were small enough ($5.3*10-96$ and $4.1*10-6$ respectively) to reject both null hypotheses. Statistically (since we rejected the null hypothesis), we can state that the YES voters posted either more or less tweets that NO users. To determine which is which (greater or less than) we used the average (other choice would be the median). Since the average number of tweets from YES voters (11.3) was greater than the average number of tweets from NO voters (11.3) we can state with high confidence (p<0.01) that: "The YES voters sent more tweets than NO voters"

Similarly, statistically, since we rejected the null hypothesis, we can state that the YES voters posted either more or less tweets containing the #VOTEYES hashtag than the number of "#VOTENO" tweets posted from NO voters. To determine which is which (greater or less than) we also used the average. Since the average number of tweets containing the "#VOTEYES" hashtag in YES users is 2.1 and the average number of tweets containing the "#VOTENO" hashtag was 1.7 we can state with high confidence (p<0.01) that: "YES users sent more #VOTEYES tweets than NO users sent "#VOTENO" tweets".

The complete source code for these calculations is available here [132].

As show in [325], it is common to deliberate make use of bots or real conscripted users to promote a particular ideology or party prior to an election event (also called "slacktivism"). In order to validate this phenomenon and measure its effect is a challenging task.

6.2.2 Entities Co-occurrence

We define co-occurring two entities when there exists at least one tweet that contains both entities. The distance between entities, as is defined as: $d = log(10 + c_{max} - c)$, where c is the number of tweets that contain a specific pair of entities, and c_{max} is the maximum c (max co-occurrence). In order to emulate the spring link attractive forces between nodes [122], we apply the "neato" visualization method of Graphviz software. Figure 6.5 shows the visualization of the distances of entity pairs with at least 500 occurrences for the referendum dataset. The distance between two entities represents the number of tweets in which they co-occur (the higher co-occurrence the closer the distance). We notice that *YES* and *NO* entities are central to the discussions, with a small in-between distance. It is also clear that Europe-related entities are closer to the *YES* point, while the entities regarding domestic affairs, including *debt* are closer to the *NO* point.

Sarcasm, Sentiment and Hashtags

Sarcasm detection revealed some interesting points concerning the use of sarcasm in the political domain. Overall, 61.8% of the total referendum tweets and 58.7% of the total election tweets had a positive sarcasm value (>0%). Nonetheless, the percentage of tweets

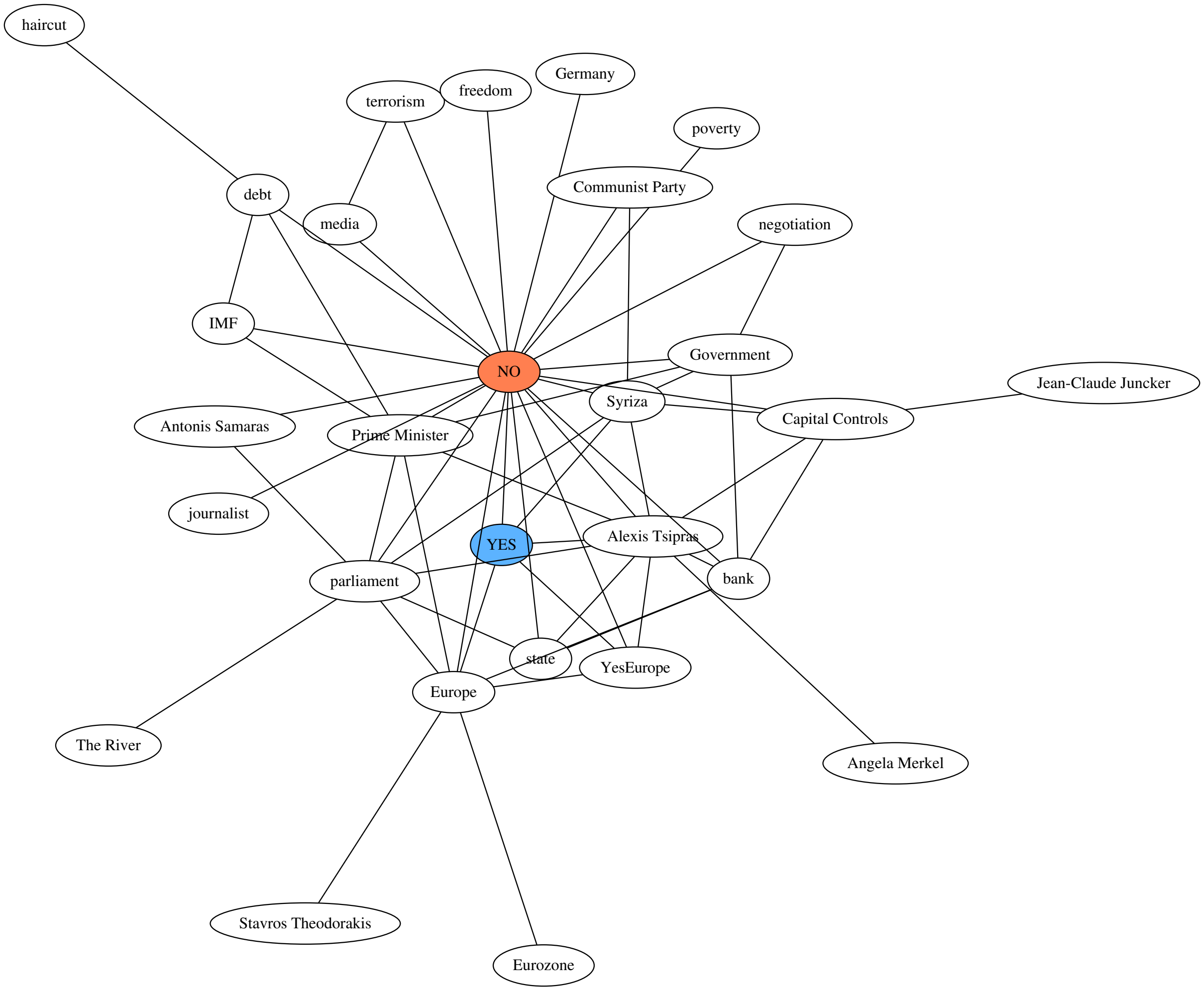


Figure 6.5: Entities co-occurrence in referendum. This graph shows the force-directed graph drawing of main entities (more than 500 tweets) of the Referendum dataset.

with strong sarcasm (>20% sarcasm value) was 27.1% and 28.8% for referendum and elections tweets, respectively. Similarly to our study, in [223] they involved sarcasm detection in 2012 US elections and found 23% sarcastic tweets. Similarly, the same percentage of 29% was also detected in a collection of tweets regarding the candidates of the Republican party, running for the US Presidential nomination for the same elections [213].

To our knowledge, the only other study though attempting to identify sarcasm in Greek political tweets, was performed in a much smaller (44,000 tweets) dataset referring to the Greek legislative elections of 2012 and concluded that 54.5% of tweets are sarcastic [73]. It is difficult to obtain a ground-truth regarding sarcasm percentages, since the subject of a tweet is a very strong indicator of sarcasm. In [158], sarcasm percentages of tweets vary from 3% to 85% according to their associated topics and "Politics" is one of the most sarcasm-prevalent ones.

Ironic posts were prevalent for specific hashtags, which, after looking into the text entities, revealed the level of the citizen aversion to the entities involved in the current situation, namely, the earlier governments and a company in the centre of talk about corruption (Figure 6.6). Contrarily, the least use of irony was found to feature the talk about the entities at stake that would be affected the most by the referendum outcome, such as Germany, Greece, Europe, and the EU.

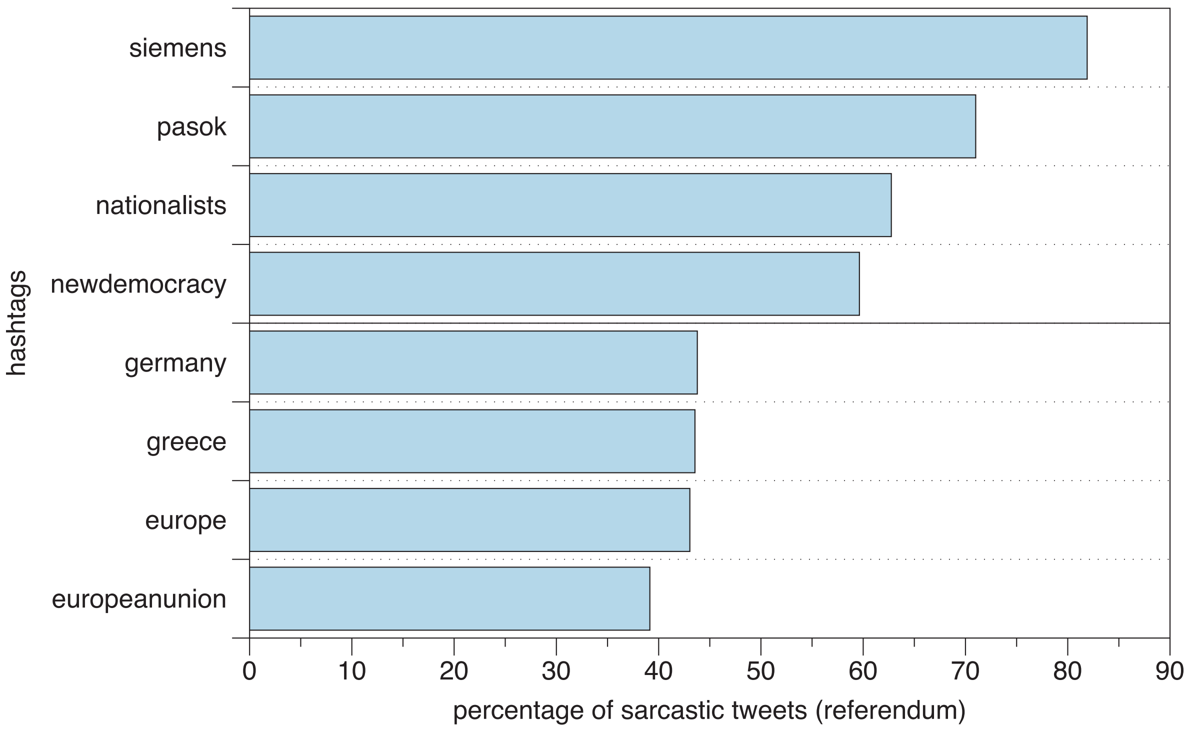


Figure 6.6: Hashtags mainly used in sarcastic and non-sarcastic posts during the pre-referendum period.

It is also worth noting that the linear regression has a negative slope, indicating a negative relation between sarcasm and number of hashtags (figure 6.7), in both referendum and elections data. Each point in these figures is a tweet. Figures (a) and (b) contain tweets in Referendum and Elections respectively. X axis contains the number of hashtags. The "sarcasm value" (y axis) ranges from -100% (definitely not sarcastic) to +100% (definitely sarcastic). Tweets with high number of hashtags exhibit lower values of sarcasm.

Despite the fact that the sarcasm assignment provided a glimpse into the thoughts of

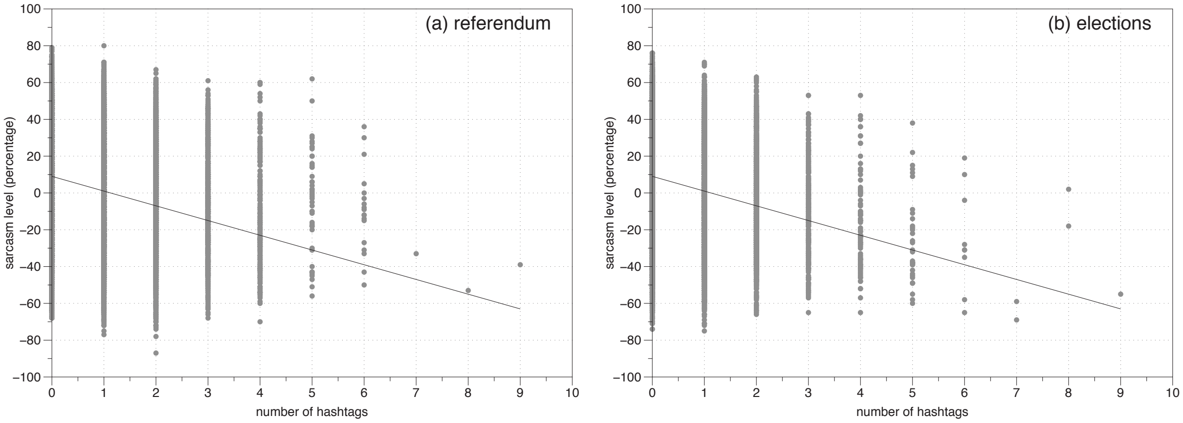


Figure 6.7: Number of hashtags and sarcasm.

the citizens revealing causes and worries related to the outcome of the referendum, there was a different but equally valuable aspect exposed by the sentiment polarity.

Figures (a) and (b) (6.8) show the entities that exhibit the highest sentiment polarization during the pre-Referendum and pre-Elections period respectively. Polarization is measured as the difference between the average positive and negative sentiment. The negative sentiment values have higher range in Referendum (from -4 to 0) than in Elections (range from -3 to 0). Sentiment values (y axis) are measured according to the SentiStrength score. It is noticeable that the citizens thought about the forces that actively tried to influence the outcome of the referendum (figure 6.8) if you look at the entities that exhibit the highest polarization of sentiment (defined as the difference between the average positive and negative sentiment values for each entity).

The tweets mentioning more than one of the highlighted entities, you see extreme polarization in those texts, clearly separating the negative sentiment towards *journalists* and the *mass media* against the positive sentiment towards *Alexis Tsipras* and *freedom*. For the elections, the same four entities exhibited the highest polarization, although new extreme positives and extreme negatives emerged (such as *terrorism* and *poverty*). The actual cause of polarization is visible by revealing the entities that exhibited the highest polarization of sentiment provided and examining the words that carried that sentiment. After the examination of the content of the referendum, the citizens perceived the journalist and mass media input as propaganda, an attempt to steer the citizens to vote for specific pro-austerity parties. It is indicated in both analyses provided, that the citizen perception of journalists and the mass media contribution to both electoral events. This was further reinforced upon examining the co-occurrence with the remaining two highly polarized entities. The positive sentiment co-occurred in the context of establishing freedom through voting for the prospective candidate, while the negative sentiment co-occurred in the context of the propaganda. Insight to the connections the citizens perceived and justified towards their elections voting was provided by co-occurred sentiment polarization.

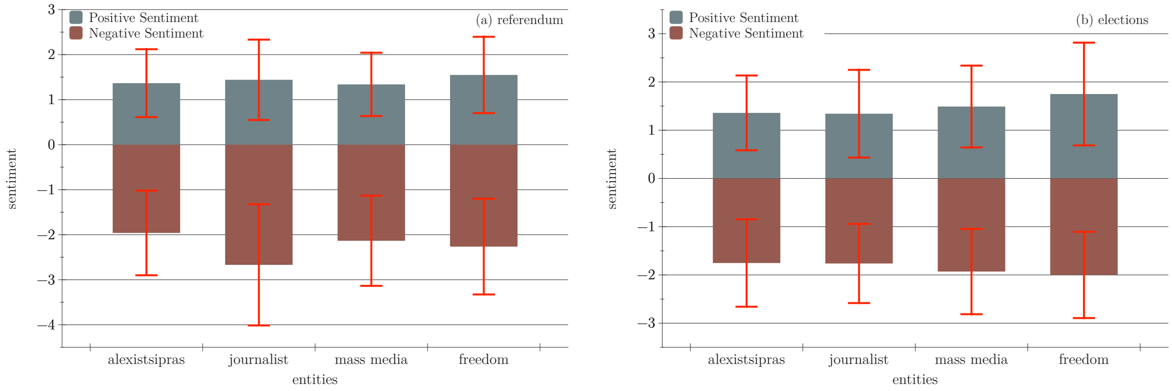


Figure 6.8: Entities with extreme sentiment polarity.

6.2.3 Temporal Variation of Sentiment

By computing the sarcasm and sentiment levels, we can visualize the temporal sentiment variation for any entity. We applied "sarcasm correction" to the sentiment for tweets with positive sarcasm to eliminate the influence of sarcasm. Specifically, each tweet sentiment was corrected towards the neutral side proportionally to the percentage of sarcasm that it contained.

Figure 6.9 shows the local linear regression lines (LOESS) of positive and negative sentiment over time for the top 5 most frequent entities of referendum and elections. The y axis showing the sentiment is encoded according to SentiStrength [10]. Positive sentiment ranges from 1 (not positive) to 5 (extremely positive). Negative sentiment ranges from -1 (not negative) to -5 (extremely negative). Here we notice that during the pre-referendum period, the positive sentiment for Europe decreases and the negative sentiment for the Greek Prime Minister Alexis Tsipras increases and becomes almost stable after the enforcement of the Capital Controls on June 29th. This trend is reversed on the elections, since the leading party, SYRIZA, undergoes a decrease in positive sentiment and an increase in negative sentiment, showing a general dissatisfaction of the party actions regarding the post-referendum political developments.

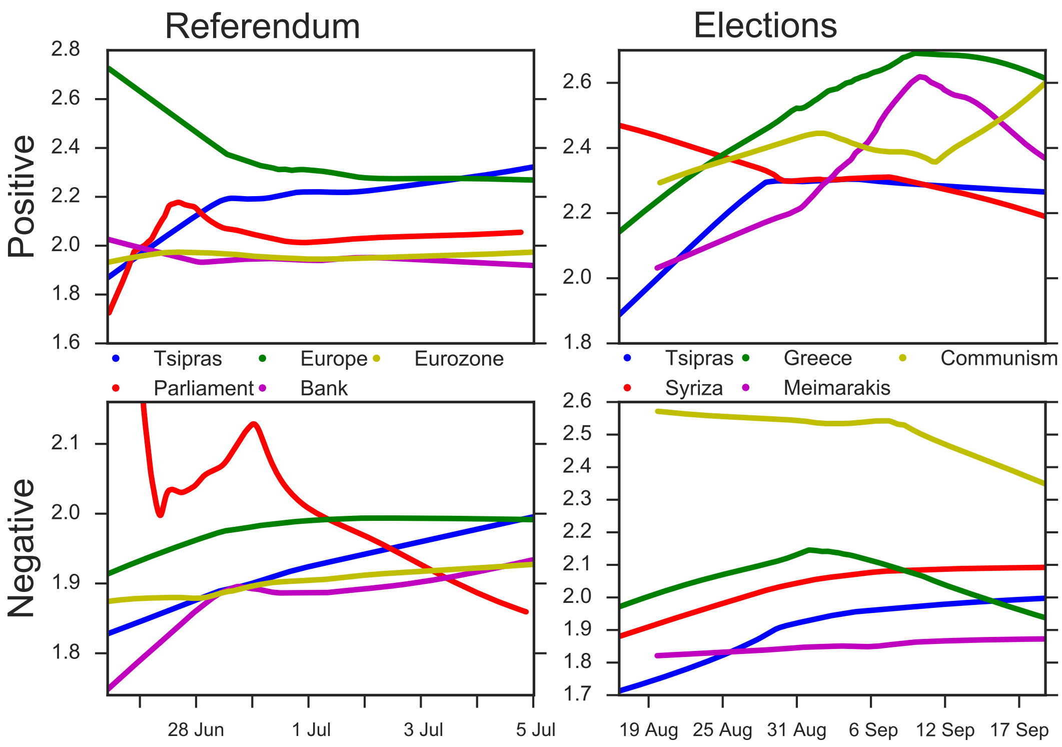


Figure 6.9: Variation of sentiment in referendum and elections.

Despite the high percentage of the "NO" vote, after the referendum, the government did the highly criticized action to accept Eurogroup's measures. We assume that may many sentiment shifts were generated for various entities, after this move. Consequently it is interesting to see how "YES" voters and "NO" voters reacted to this development. So we split users into two disjoint groups: the "YES" voters and the "NO" voters, in order to measure this and we kept only users that have posted in both referendum and elections datasets. We measured the average positive and negative sentiment for each entity in both datasets, for each user in every group. Finally, we applied the Mann-Whitney U test between the average sentiment of this group in the referendum dataset and the average sentiment of this group in the elections dataset, for each entity, user group and sentiment. Figure 6.10 shows the entities for which the sentiment was significantly changed ($p < 0.001$). The arrows show statistically significant changes in the average sentiment between referendum and elections for the same group of users ("YES" voters and "NO" voters). Negative values represent negative sentiment, while positive values represent positive sentiment. A greater absolute value for a negative or positive sentiment, signifies that the sentiment is more intense. The vectors of this change are portrayed by the direction and the length of the arrows in the figure. The same figure shows a general shift of negative sentiments towards neutrality (the only exception is the "Debt" entity for the "NO" voters). Indeed, elections (which constitute a more frequent electoral event), did not attract the same negatively charged content. Also "YES" voters expressed more positive comments regarding "ND" (i.e., the main opposition party) and the prime minister Alexis Tsipras, after the referendum.

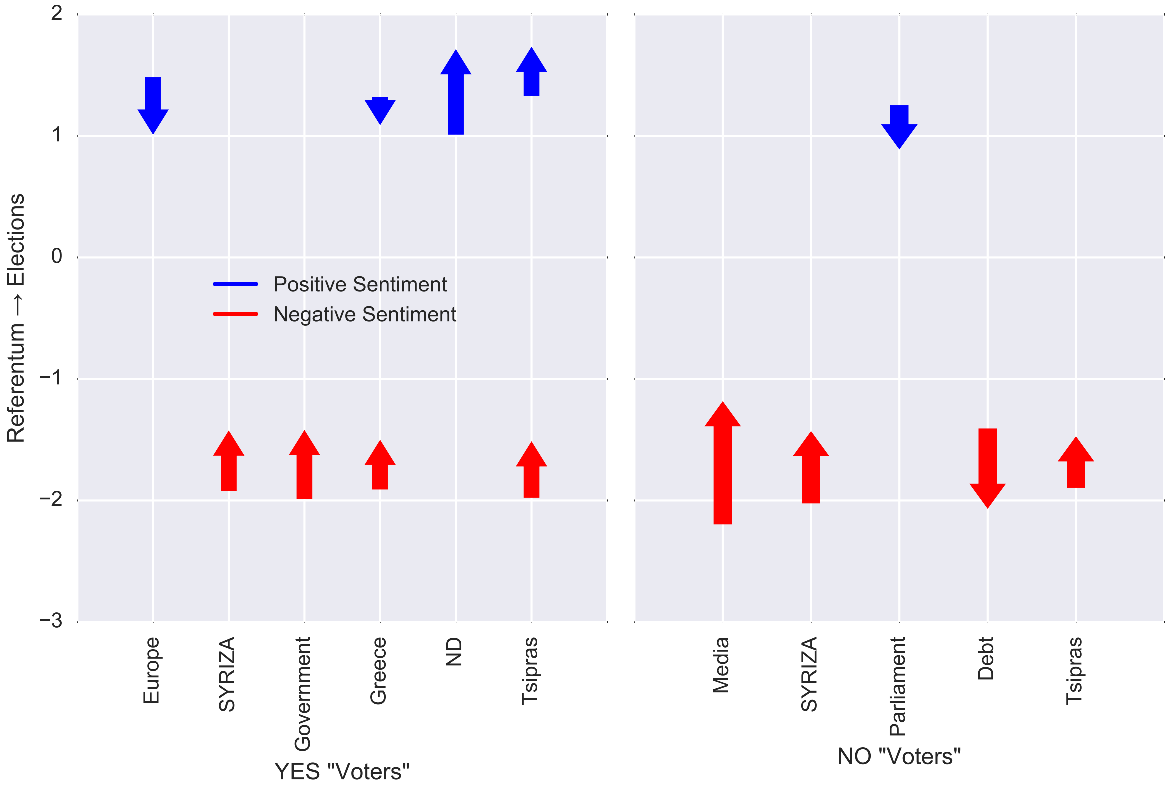


Figure 6.10: Change of sentiment between Referendum and Elections..

6.2.4 Topic Modeling

Topic modeling consists a powerful tool to detect thematic patterns in text corpus. Interesting ideological inclinations, tendencies and concept proximities can be revealed if it is applied in political discussions. One of the most common method used in this area is Latent Dirichlet Allocation (LDA), which has been used in the past to analyze online content, like news items and blog posts, and for spam detection.

Generally, topic modeling generates a predetermined number of topics. A per-entity distribution, or else the probability that an entity belongs to a topic is computed for each topic. Topics can be also projected in a 2-dimensional space for better visualization. The results gives two topics lying in distant places after LDA analysis, indicating that they have very different mixture of entity probabilities. In contrast two proximal topics indicate a concordance of entity probabilities.

Background work shows that this method has also been used to analyze political content in Twitter. LDA was used to quickly identify emerging topics [261] in Twitter, during the German federal elections of 2013. This technique was able to detect prevalent discussion topics earlier than Google Trends. In different work they analyze the extracted topics from tweets regarding Barack Obama [215]. In order to locate the most insightful opinions in each topic, the authors, applied content summarizations methods. Our work targets to study the semantic distance between prevalent opponent entities in both Referendum and Elections. Topic modeling reveals the entities that exhibit semantic similarity, while sentiment analysis reveals the overall positive and negative emotions that characterize each entity.

The application of LDA in political-related Twitter content is challenging [192] mainly because of the short length of Twitter posts, the special linguistic elements that they contain and the variability of the political discussion, We propose performing entity identification combined with sarcasm filtering, we can efficiently locate dominant topics in Twitter.

In our work after excluding all tweets that had sarcasm identifier value higher than 5%, we analyzed the manually-identified entities in the tweets with Gensim Python library. We used LDAvis [279] that performs a Principal Component Analysis (PCA) to project the identified topics on the 2-dimensional space in order to visualize the generated topics. Additionally we measured the average positive and negative sentiment across all tweets, for each entity. Each topic contains: (i) a set of entities, and (ii) the proportion by which an entity belongs to a topic, e.g. entity *Greece* might belong by 70% to topic 1 and by 30% to topic 2. The LDA topic analysis for the referendum and the elections are shown in figures 6.11 and 6.12 respectively.

Each circle in figures 6.11, 6.11 is a topic placed according to PCA. Circle size is proportional to the marginal distribution of each topic. The 4 most frequent entities, that each topic contains, and their average positive sentiment (blue bars) and negative sentiment (red bars) are also shown.

Fig 6.11 shows the topics containing *VoteYes* and *VoteNo*, which seem to be unexpectedly close. Entities associated with the *VoteNo* topic contain stronger positive sentiments (except the *Conservatives* entity, while entities associated with the *VoteYes* topic contain stronger negative sentiments. Interestingly, despite the fact the Prime Minister (Tsipras) and his party (SYRIZA) were strong "NO" supporters, the topic that contains them lie in the middle of the "YES" and "NO" topics. It is worth mentioning a topic on the left of the figure with prevalent negative sentiments that contain entities pervasive to the anti-austerity discourse like "Varoufakis", "Troika" and "Eurogroup".

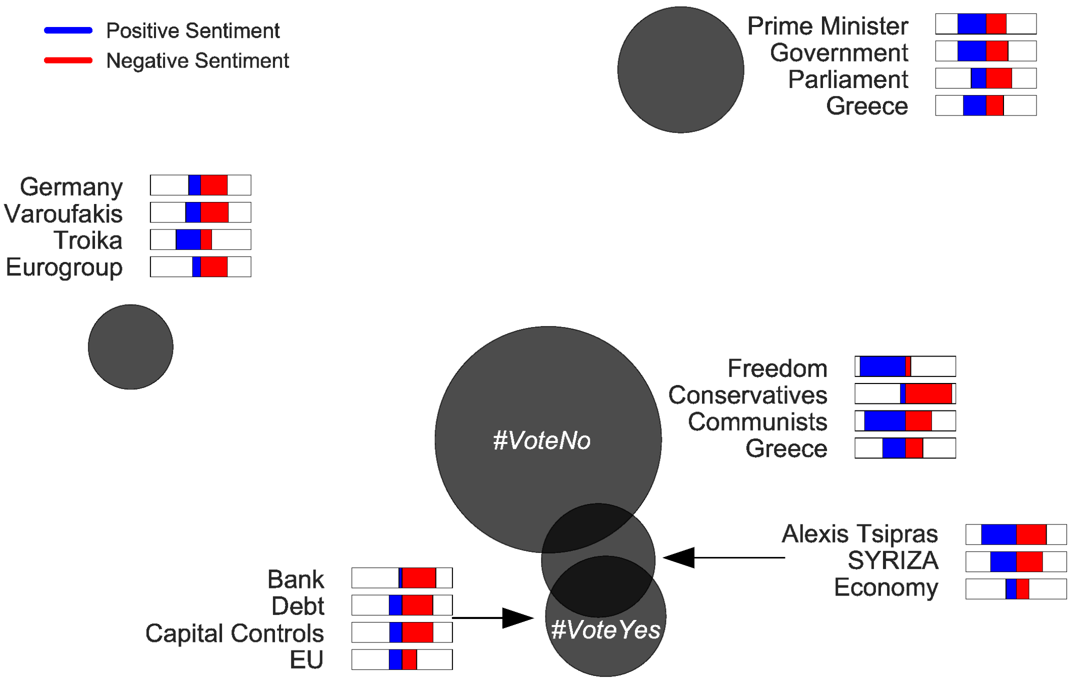


Figure 6.11: LDA Topic Model of Referendum Entities.

In figure 6.12 we observe that topics regarding elections are placed according to the political spectrum Conservative parties are represented by the two conjoined circles at the left, while the rest three circles represent (i) the dominant center-left *SYRIZA* party at the top, (ii) politically center entities (namely, *PASOK*) in the middle, and (iii) the far-left parties at the bottom. We assume that this placing is because the conservative parties in Greece were more strongly affiliated than the left and center-left parties.

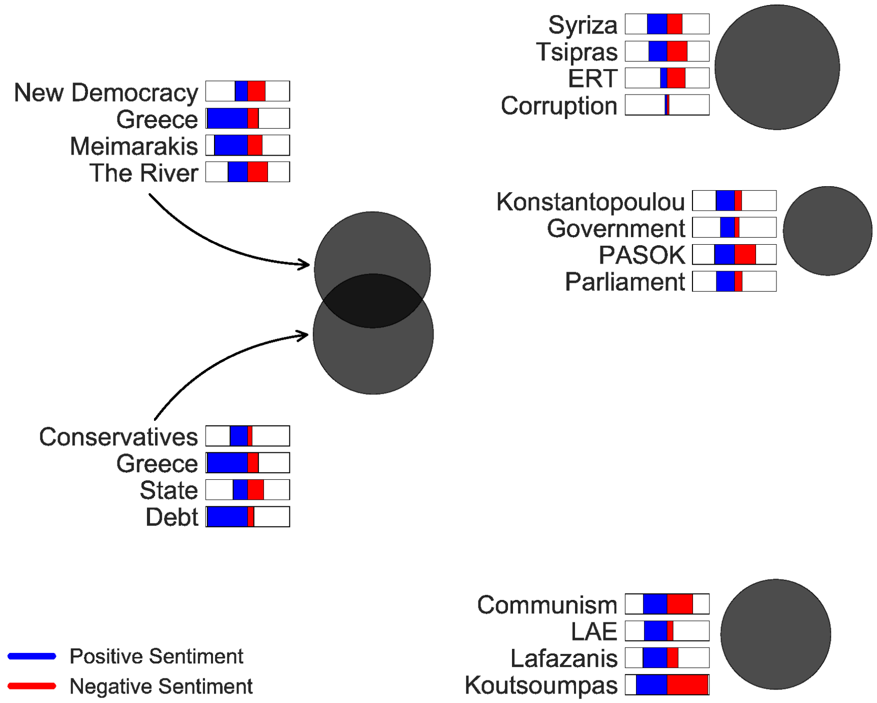


Figure 6.12: LDA Topic Model of Elections Entities.

Chapter 7
Conclusion

7.1 Synopsis of Contributions

In this thesis we present a multitude of methods for the analysis of Twitter, a popular and vibrant Online Social Network. The analysis is threefold; initially, we study the average degree and the temporal growth of Twitter, in order to understand the growth dynamics of the underlying social graph. Next, we apply machine learning techniques for the study of the malicious nature of tweets spreading through trending topics, forming spam campaigns and reveal their inner structure through graph analysis. Finally, we perform a sentiment analysis and topic modeling in a political dataset in Twitter.

These three studies cover a complete analysis of the Twitter data, taking into account some of the major points of views in Twitter science. We acknowledge the fact that there are aspects of research in Twitter that are very active and we have not covered. These areas include bot detection, sybil attacks and privacy attacks that attempt to reveal private user information like the real name, location, gender and age of the users. Also, a very important area of research is fake-news detection and the estimation of their impact in real social events, like political campaigns. We are confident that the methods presented in this thesis can be extended for the analysis of these phenomena and can also be used to build effective defence mechanisms that will result on limiting their negative effects in society.

In the following paragraphs we comment on the major findings of this thesis and point out directions for future work and improvements.

7.1.1 The temporal evolution of Twitter

The models of the evolution of Social Networks are extremely important for elucidating their structure and explaining the behaviour of their users. However, the task of modeling can be really challenging, due to the enormous size of modern OSNs and the prohibitive computational complexity of many essential graph properties. This thesis, initially, shows a computationally efficient method to model the growth of an OSN, based on the simple property of node average degree. Our techniques is divided in four distinct parts:

- The application of the heuristic that approximates the friendship time-creation, with a time complexity of $O(|E|)$.
- The sorting of all edges according to this approximation.
- The calculation of the average outdegree of the network for every day between June 2006 and January 2015 (in total 3.100 days). The time complexity of this part is $O(T|V|)$ where T is the total number of time periods ($T = 3.100$).
- Fitting the average outdegree for all days to the "Leskovec model", which requires minutes of computation in a commodity computer.

Sorting all edges of the network is the computational challenging part, which can be easily parallelized. The rest part of the computation can take place in a single workstation. Overall, the complete computation required approximately one day, in a high end workstation (single 4-core Intel i7 processor, 3.4GHz, 16Gb RAM). The experiments that took place in this study show approximately the same growth periods that could be delineated from two fundamentally different samples of Twitter. The first (*BIG*) dataset contains 92 million users and was created with the Random Walk sampling method. The second (*KWAK*) dataset is approximately half of the size of the first and contains the friends and followers of a relatively old (2009) dataset.

Additionally, we show how this method can portray fluctuations of growth even years before the sampling of the OSN happened. Particularly, we focus on events that happened more than 5 years before the OSN was sampled. There is an open question whether the moments of the outdegree distribution of Twitter are well defined, although it is more likely not be a power-law. The mean of a power-law distribution, with exponent $\lambda < 2$ diverges, meaning that repetitive measurement in independent samples will result in very large fluctuations [236]. The latest and largest study [230] concluded that this distribution is best fit by a log-normal function, which has all its moments well defined.

In order to fit the "Leskovec model", we measured the average degree of the same sample of the social graph on a day-by-day base, instead of measuring the average outdegree in independent samples, therefore we did not expect large fluctuations. This is also shown in figure 4.1, where the plot has a relative smooth form of the average outdegree. However, the proper elucidation of the form of this distribution with adequately large sample sizes is a crucial open question that needs to be addressed.

In order to establish reliable causal relationships between the presented events and the alterations of the growth exponent, future work is needed. We include in our future plans to apply heuristics for the estimation of the second half of the Leskovec model parameter, which is the diameter allowing us to estimate the effect and size of the "Shrinking parameter" of Twitter. Finally, we believe that this approach will help researchers to model efficiently the evolution of large OSNs and delve into their past in order to investigate "which" and more important "how" specific events alter their growth patterns.

Future Work. Some extra future steps are needed in order to establish reliable causal relationships between the presented events and the alterations of the growth exponent. We plan to apply heuristics for the estimation of the second half of the "Leskovec" model parameter, which is the diameter, allowing us to estimate the effect and size of the "Shrinking parameter" of Twitter. Finally, we believe that this approach will help researchers to model efficiently the evolution of large OSNs and delve into their past in order to investigate "which" and more important "how" specific events alter their growth patterns.

7.1.2 Identifying spam campaigns in Twitter

Having studied the growth patters of Twitter through the average degree, we conducted a study regarding the characteristics of spam propagating through Twitter. Specifically, we look how spammers try to increase the effectiveness of their campaigns by using certain features of the service. The study of our 3-month dataset revealed a set of spam campaigns that exhibited a stealthier approach than other campaigns and also masqueraded URLs pointing to spam websites as a Google search result. We used a set of 86 blacklists (RBLs) and a heuristic for identifying these new campaigns, we created labelled datasets that we used for training a classifier.

Next we quantified the behaviour of legitimate users and spammers, using metrics for Twitter-specific keywords and content and discovered a large divergence in categories such as the number of trending topics included in tweets. Additionally, we build a classifier for a spam detection mechanism that uses these metrics and tested it on a dataset containing all the tweets collected during a period of one month. The detection mechanism can correctly identify 73.5% of the stealthy spammers, while maintaining a very low false positive ratio of 0.25%. Overall, our system offers a light detection mechanism for a stealthy and persistent class of Twitter campaigns, while maintaining a very low false positive rate that is a very significant requirement for deployment in a real environment and relies on features that require low to zero computational resources since they are widely available from Twitter's API.

Future Work. Concerning our classifier in the spam domain, there is an open question whether a classifier trained in one month dataset, can accurately classify Twitter users of another or whether the training needs to be repeated. We plan to validate this on our future work. Also, we intend to inspect closer the nature of the feature of the tweets, hoping that we will get insights that might increase the accuracy of the classifier.

Our analysis does not make any distinction between hijacked accounts posting malicious content in Twitter and dedicated spam accounts could shed more light in this. We plan to explore this area in the future.

The nodes of a spam domain has an average edge degree of 125, which means that every spam domain in the graph could be found in average, in the tweets of 125 different users. However, the average edge degree for the graph depicting users and the URLs they tweeted for that day is 2.3 for legitimate URLs.

Finally, we plan to explore diverse graph properties from the spam campaign graphs, or the user-URL graphs, in order to identify spam users and campaigns and study other ways that spammers use to obfuscate their malicious links like goggle search results, found in this part of the thesis.

7.1.3 Sentiment and topic analysis in Twitter

Next, we studied two Twitter datasets from two politically associated electoral events and applied sentiment analysis and sarcasm detection. One of the steps was the entity detection which combined manual, semi-automatic and scripted processing, as well as lexical resources to correctly assign sentiment. This combination was necessary for tackling the traditionally hard-to-analyze political domain, by blending entity-level sentiment and data statistics. The results revealed societal and political trends that are commonly unnoticed, that guide citizen choices and actions, which traditional polls fail to detect. The included exploratory analysis shed light into part of the public sentiment towards main entities along with their semantic proximities and was applied to two related electoral events, enabling the creation of lexical resources that covered the semantic content of a wide and complex political background. As an extension of the previously existing resources, these resources, were used for entity and sentiment detection and are available both as general-purpose resources and most importantly, optimized for the political domain.

Most of the available studies in social networks regarding electoral or political events, focus on specific aspects of the data (e.g., sentiment analysis, entity identification). We propose in this part of the thesis that in-depth discovery in similar datasets should include at least 5 types of analysis: volume analysis, entity identification, sentiment detection, sarcasm correction and topic analysis. The results of an analysis like that will improve other types, i.e. sentiment analysis enhancements can improve topic analysis. The main target of this part of the dissertation is to reveal quantitative aspects of data analytics that may be proven helpful for political analysts and citizens alike. The semantic interpretations of the results may be suited accordingly by the interested parties, making social interpretations on the qualitative level. Consequently, although these is a distance from creating qualitative conclusions about the underlying social dynamics affecting a political discourse, we can absolutely be assisted towards that goal by making sense of the vast amounts of social data through this approach.

Future Work. The results of this analysis of this thesis certainly hinted further work. A very interesting next step for better understanding citizens and society, could be to detect emotion (sadness, happiness, fear, anger, etc.) and see how emotion drives societal and consequently, political changes, since sentiment is a descriptive work for all emotions and not all emotions are the same [111]. Additionally, our future plans include:

- Sentiment consistency (is there a specific sentimental spectrum for each user?);
- Context-specific opinions (does sentiment give insights regarding opinions on certain entities?);
- Sentiments at the phrase or expression level (instead of per-word sentiment assignment, can we assign sentiment to the whole sentence by incorporating contextual subjectivity information?).

Also a missing part from our analysis that needs to be implemented are techniques to detect bots that are massively employed to spread content, in favor of a specific candidate. This is a common phenomenon in many elections, where some examples are from the U.S presidential elections in 2016 [167] and the 2013 Australian Federal Election [325]), with yet unknown impact on the election results. We do not have any other supporting indications that this phenomenon took place on the events that we analyzed, except from the significantly higher number of average tweets posted by "YES" voters in the referendum dataset. We plan to analyze future electoral events mainly in the European area, in order to advance our technique towards extracting qualitative insights regarding users' political affiliations and reveal potential malicious efforts to obscure the online discourse.

7.2 Open research questions in Twitter Science

In this last section of the thesis we would like to discuss some wider open questions that remain mainly unsolved in Twitter. First and foremost is a policy issue and has to do with data from Twitter. Given the current limitations of Twitter's API, it is practically impossible to collect enough data to confidently measure the network properties of the social graph. Also, Twitter's Terms of Service, explicitly forbids the mass collection and sharing of data. Therefore, areas like spam detection, bot characterization, sentiment analysis and topic identification suffer from the lack of open dataset and golden standards that could enhance the reproducibility of studies and could ease comparison efforts. Finally, Twitter's API restricts significantly the access to past data with enormous societal and historical value. To alleviate this problem, researchers and policy makers could form alliances that could inform the public for this issue and force Twitter to change its policy. Of course this constitutes a distinct research question. Perhaps a sophisticated data collection approach could overcome in the future these limitation without violating Twitter's Terms of Service.

Concerning the social graph itself, the elucidation of an accurate and predictive model that describes the evolution of Twitter is still an open question. Similarly, the relationship between the Twitter's social network and the underlying real social network is unclear. It would be very interesting for example if we could differentiate the accounts that we treat as news providers and the accounts that we consider as real-life "friends". We also located that some of the most basic properties of the social graph have not yet measured undoubtedly, or the studies that have measured them are more than five years old. For example, one of the most basic graph properties, the node degree distribution is not yet known. Studies are osculating between a power-low to a log-normal distribution, where different studies give different parameters of these distributions. Another property is the diameter. Given the computational complexity of this attribute, we can only approximate its value with heuristics. The proper characterization of these values for various time-points of the Twitter's social graph will give a definite answer to the evolution nature of Twitter.

Concerning sentiment analysis, we notice that most, if not all of existing studies perform analysis on an ad-hoc sample of Twitter, collecting all tweets that contain a set of hashtags. It is unclear if the data collected are enough to capture a sentiment towards an entity. A study that could assess different sampling techniques for assessing different types of sentiment could be very useful for this purpose. Additionally, Twitter, as a worldwide OSN, is able to capture the sentiment towards various entities from different cultures and languages. These studies do not perform comparative analysis, although there are many studies that perform multi-lingual sentiment analysis. Namely, they do not compare the public sentiment towards the same entity. Shedding some light on the differential ways on which different cultures perceive the same entity (for example the president of USA, immigration, climate change, etc.) would be more than interesting.

Concerning spam classification and bot detection, there is some excellent research performing very sophisticated method for these tasks. Yet, all these studies pinpoint the need for *timely* detection. Time is of extreme essence regarding the success of spam campaigns and the spread of fake news. Therefore, we believe that spam and bot detection methods should make time a first priority attribute in efficiency measurements. Subsequently, browser plugins that could indicate spam, bot and fake-news tweets in real-time could battle efficiently these attacks.

We hope that future research, with the assist of the methods presented in this thesis, will deal soon with these issues with effective and open solutions.

Bibliography

[1] China blocks twitter, flickr and hotmail ahead of tiananmen anniversary. https://www.theguardian.com/technology/2009/jun/02/twitter-china". [Online; Accessed on 11/10/2019].

[2] Greek Referendum 2015 demographics. http://www.publicissue.gr/en/2837/. [Online; Accessed on 1 August 2019].

[3] The sarcasm detector. http://www.thesarcasmdetector.com/. [Online; Accessed on 02/25/2019].

[4] Twitter blows up at sxsw conference. http://gawker.com/243634/twitter-blows-up-at-sxsw-conference". [Online; Accessed on 11/10/2019].

[5] Why The Polls In Greece Got It Wrong? https://www.huffpost.com/entry/greece-polls-wrong_n_7754874. [Online; Accessed on 1 August 2019].

[6] Making The Switch From Twitter to Jaiku. http://goo.gl/JMuhKA, 5 2007. [Online; accessed 10-October-2015].

[7] Twitter Survives Stevenote - But FriendFeed was the Place To Be. http://goo.gl/aGyGW0, 6 2008. [Online; accessed 10-October-2015].

[8] State of Twitter Spam. https://blog.twitter.com/2010/state-of-twitter-spam, 3 2010. [Online; accessed 10-October-2015].

[9] The March of Twitter: Analysis of How and Where Twitter Spread. https://goo.gl/RiWs4n, 8 2010. [Online; accessed 10-October-2015].

[10] SentiStrength - sentiment strength detection in short texts - sentiment analysis, opinion mining. http://sentistrength.wlv.ac.uk/, 2017. [Online; Accessed on 1 August 2019].

[11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended semantic web conference - ESWC'11*, pages 375–389. Springer, 2011.

[12] Anupama Aggarwal, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. PhishAri: Automatic realtime phishing detection on Twitter . In *eCrime Researchers Summit '12*, pages 1–12. IEEE, oct 2012.

[13] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the International Conference on Social Media and Society - SMSociety'18*, pages 226–230. ACM, 2018.

[14] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. volume 401, pages 130–131. Nature Publishing Group, 1999.

[15] S. Aleksandar. How Much Time Do People Spend on Social Media in 2019? https://bit.ly/2kyGC1w. [Online; Accessed 2019-09-22].

[16] Alexa Internet, Inc. Alexa Top 500 Global Sites. http://www.alexa.com/topsites. [Online; Accessed: 2018-10-28].

[17] Abdullah Almaatouq, Ahmad Alabdulkareem, Mariam Nouh, Erez Shmueli, Mansour Alsaleh, Vivek K Singh, Abdulrahman Alarifi, Anas Alfaris, and Alex Sandy Pentland. Twitter: who gets caught? observed trends in social micro-blogging spam. In *Proceedings of the ACM conference on Web science - Websci'14*, pages 33–41. ACM, 2014.

[18] Juan Pablo Alperin, Erik Warren Hanson, Kenneth Shores, and Stefanie Haustein. Twitter bot surveys: A discrete choice experiment to increase response rates. In *Proceedings of the International Conference on Social Media and Society - SMSociety'17*, #SMSociety17, pages 27:1–27:4, New York, NY, USA, 2017. ACM.

[19] Mansour Alsaleh, Abdulrahman Alarifi, Abdul Malik Al-Salman, Mohammed Alfayez, and Abdulmajeed Almuhaysin. Tsd: Detecting sybil accounts in twitter. In *International Conference on Machine Learning and Applications - ICMLA'14*, pages 463–469. IEEE, 2014.

[20] Amit A. Amleshwaram, A. L. Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang. Cats: Characterizing automation of twitter spammers. In *International Conference on Communication Systems and networks - COMSNETS'13*, pages 1–10. IEEE, 2013.

[21] Amnesty International. Troll patrol findings, using crowdsourcing, data science & machine learning to measure violence and abuse against women on twitter. https://bit.ly/2QAQZk9, 2018. [Online; Accessed 2018-12-30].

[22] Paul André, Michael Bernstein, and Kurt Luther. Who Gives a Tweet?: Evaluating Microblog Content Value. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing - CSCW '12*, CSCW '12, New York, NY, USA, 2012. ACM.

[23] ANT1 DIGITAL Ant1news. Metron Analysis for the referendum. https://www.ant1news.gr/General/article/414803/mprosta-to-oxi-stin-teleytaia-dimoskopisi-tis-metron-analysis. [Online; Accessed on 5 November 2019].

[24] Antonio Fernández Anta, Luis Núnez Chiroque, Philippe Morere, and Agustín Santos. Sentiment analysis and topic detection of spanish tweets: A comparative study of of nlp techniques. *Procesamiento del lenguaje natural - sepln'13*, 50:45–52, 2013.

[25] Despoina Antonakaki, Sotiris Ioannidis, and Paraskevi Fragopoulou. Utilizing the average node degree to assess the temporal growth rate of twitter. *Social Network Analysis and Mining - SNAM'18*, 8(1):12, 2018.

[26] Despoina Antonakaki, Iasonas Polakis, Elias Athanasopoulos, Sotiris Ioannidis, and Paraskevi Fragopoulou. Think before rt: An experimental study of abusing twitter trends. In *International Conference on Social Informatics - Socinfo-'14*, pages 402–413. Springer, 2014.

[27] Despoina Antonakaki, Iasonas Polakis, Elias Athanasopoulos, Sotiris Ioannidis, and Paraskevi Fragopoulou. Exploiting abused trending topics to identify spam campaigns in twitter. *Social Network Analysis and Mining - SNAM'16*, 6(1):48, 2016.

[28] Despoina Antonakaki, Iasonas Polakis, Elias Athanasopoulos, Sotiris Ioannidis, and Paraskevi Fragopoulou. Exploiting abused trending topics to identify spam campaigns in twitter. *Social Network Analysis and Mining-SNAM'16*, 6(1):48, 2016.

[29] Despoina Antonakaki, Dimitris Spiliotopoulos, Christos V Samaras, Sotiris Ioannidis, and Paraskevi Fragopoulou. Investigating the complete corpus of referendum and elections tweets. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining -ASONAM'16*, pages 100–105. IEEE, 2016.

[30] Despoina Antonakaki, Dimitris Spiliotopoulos, Christos V Samaras, Polyvios Pratikakis, Sotiris Ioannidis, and Paraskevi Fragopoulou. Social media analysis during political turbulence. *PloS one*, 12(10):e0186836, 2017.

[31] Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors. *Semantic Enrichment of Twitter*

Posts for User Profile Construction on the Social Web, pages 375–389. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[32] Salman Aslam. Twitter by the numbers: Stats, demographics & fun facts. https://www.omnicoreagency.com/twitter-statistics/. [Onlne; Accessed: 2018-10-27].

[33] Sitaram Asur and Bernardo A. Huberman. Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499. IEEE, aug 2010.

[34] Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, and Chunyan Wang. Trends in Social Media: Persistence and Decay. *SSRN Electronic Journal*, 2011.

[35] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: a case study. In *the International Conference on Recent Advances in Natural Language Processing - RANLP'05*, Borovets, BG, 2005.

[36] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *International Conference on Language Resources and Evaluation -Lrec'10*, volume 10, pages 2200–2204, 2010.

[37] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, pages 33–42, New York, New York, USA, June 2012. ACM Press.

[38] David A. Bader, Shiva Kintali, Kamesh Madduri, and Milena Mihail. Approximating Betweenness Centrality. In *Algorithms and Models for the Web-Graph*, pages 124–137. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[39] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer. In *Proceedings of the ACM international conference on Web search and data mining - WSDM '11*, page 65, New York, New York, USA, 2011. ACM Press.

[40] Alexandra Balahur and Marco Turchi. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing - RANLP'13*, pages 49–55, 2013.

[41] A. Barabási. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, oct 1999.

[42] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Cascade-based community detection. In *Proceedings of the ACM international conference on Web search and data mining - WSDM'13*, pages 33–42. ACM, 2013.

[43] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Who to follow and why: link prediction with explanations. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'14*, pages 1266–1275. ACM, 2014.

[44] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

[45] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, 2015.

[46] Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2010.

[47] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[48] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[49] Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, and Peter Sheridan Dodds. An evolutionary algorithm approach to link prediction in dynamic social networks. *CoRR*, abs/1304.6257, 2013.

[50] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science - JCS'12*, 3(5):388–397, 2012.

[51] Johan Bollen, Huina Mao, and Alberto Pepe. Determining the public mood state by analysis of microblogging posts. In *In Proceedings of the Alife XII Conference*, 2010.

[52] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the International AAAI Conference on Weblogs and Social Media- ICWSM'11*, 11:450–453, 2011.

[53] Erik Borra and Bernhard Rieder. Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management - AsLib'14*, 66(3):262–278, 2014.

[54] Matko Bošnjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho: a distributed focused crawler to support open research

with twitter data. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1233–1240. ACM, 2012.

[55] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Hawaii international conference on System sciences - hicss'10*, pages 1–10. IEEE, 2010.

[56] Peter Bray. Social Authority: Our Measure of Twitter Influence. http://moz.com/blog/social-authority. [Online; accessed 10-October-2015].

[57] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[58] A. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society.

[59] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[60] David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384, 2018. PMID: 30138075.

[61] David A Broniatowski, Michael J Paul, and Mark Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

[62] Martin Bryant. Twitter Geo-fail? Only 0.23% of tweets geotagged. https://thenextweb.com/2010/01/15/twitter-geofail-023-tweets-geotagged/, 2010.

[63] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233, 2016.

[64] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[65] BRYAN BURROUGH. FYRE FESTIVAL: ANATOMY OF A MILLENNIAL MARKETING FIASCO WAITING TO HAPPEN. https://www.vanityfair.com/news/2017/06/fyre-festival-billy-mcfarland-millennial-marketing-fiasco, 2017.

[66] Nanette Byrnes. How the Bot-y Politic Influenced This Election. https://bit.ly/2fBN13R, 2016. [Online; Technologyreview.com, Accessed: 2018-12-30].

[67] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the international workshop on multimedia data mining - MDMKDD'10*, page 4. ACM, 2010.

[68] 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 1–10, New York, New York, USA, 2010. ACM Press.

[69] Daniele Cenni, Paolo Nesi, Gianni Pantaleo, and Imad Zaza. Twitter vigilance: a multi-user platform for cross-domain twitter data analytics, nlp and sentiment analysis. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[70] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'11*, 2010.

[71] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the international conference on World wide web - WWW '09*, page 721, New York, New York, USA, 2009. ACM Press.

[72] Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. Detecting irony on greek political tweets: A text mining approach. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, EANN '15, pages 17:1–17:5, New York, NY, USA, 2015. ACM.

[73] Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence*, 51:50–57, 2016.

[74] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *ICDM*, pages 817–822, 2016.

[75] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. *arXiv preprint arXiv:1807.07961*, 2018.

[76] Na Cheng, R. Chandramouli, and K. P. Subbalakshmi. Author gender identification from text. *Digit. Investig.*, 8(1):78–88, 2011.

[77] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selcuk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'10*, pages 34–41, 2010.

[78] Abdur Chowdhury. State of Twitter Spam. https://bit.ly/2QwmB5G, 2010. [Online; Twitter.com, Accessed: 2018-12-30].

[79] Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis. Entity resolution in the web of data. *Synthesis Lectures on the Semantic Web*, 5(3):1–122, 2015.

[80] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[81] Zi Chu, Indra Widjaja, and Haining Wang. Detecting social spam campaigns on twitter. In *International Conference on Applied Cryptography and Network Security*, pages 455–472. Springer, 2012.

[82] Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction. In *Proceedings of ACM SIGCOMM conference on Internet measurement - IMC '08*, page 57, New York, New York, USA, 2008. ACM Press.

[83] Jessica Elan Chung and Eni Mustafaraj. Can collective sentiment expressed on twitter predict political elections? In *AAAI*, volume 11, pages 1770–1771, 2011.

[84] Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on twitter : an unsolicited public opinion poll. *PloS one*, 10(8):e0136092, 2015.

[85] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[86] nicholas Confessore, Gabriel J.X. Dance, Richard Harris, and Mark Hansen. The follower factory. https://nyti.ms/2rJ8YZM. [Online; Accessed: 2018-10-20].

[87] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Political polarization on twitter. 2011.

[88] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter . *Proceedings of the International AAAI Conference on Weblogs and Social Media- ICWSM'11*, 133:89–96, 2011.

[89] Josh Constine. Facebook sees 2 billion searches per day, but it's attacking Twitter not Google. https://tcrn.ch/2aL3jGk, 2016. [Online; Accessed: 2018-12-30].

[90] Gordon V. Cormack. Email Spam Filtering: A Systematic Review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, apr 2008.

[91] Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Asia information retrieval symposium*, pages 238–249. Springer, 2011.

[92] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[93] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[94] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

[95] Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing - RANLP 2013*, pages 198–206, 2013.

[96] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[97] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.

[98] Thomas Dimson. Emojineering part 1: Machine learning for emoji trends. https://bit.ly/2PcHKBm. [Online; Accessed: 2018-10-7].

[99] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on web search and data mining - WSDM'08*, pages 231–240. ACM, 2008.

[100] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM, 2015.

[101] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

[102] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the international conference on World wide web - WWW'10*, pages 331–340. ACM, 2010.

[103] Rehab M Duwairi, Raed Marji, Narmeen Sha'ban, and Sally Rushaidat. Sentiment analysis in arabic tweets. In *International conference on Information and communication systems - icics'14*, pages 1–6. IEEE, 2014.

[104] Nugroho Dwi Prasetyo and Claudia Hauff. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, HT '15, pages 149–158, New York, NY, USA, 2015. ACM.

[105] Fabon Dzogang, Stafford Lightman, and Nello Cristianini. Diurnal variations of psychometric indicators in twitter content. *PLOS ONE*, 13(6):1–18, 06 2018.

[106] David Ediger, Karl Jiang, Jason Riedy, David A. Bader, and Courtney Corley. Massive Social Network Analysis: Mining Twitter for Social Good. In *2010 39th International Conference on Parallel Processing*, pages 583–593. IEEE, sep 2010.

[107] Chad Edwards, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton. Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33:372–376, 2014.

[108] Miles Efron. Hashtag Retrieval in a Microblogging Environment. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 787–788, New York, NY, USA, 2010. ACM.

[109] Young-Ho Eom, Michelangelo Puliga, Jasmina Smailovic, Igor Mozetic, and Guido Caldarelli. Twitter-based analysis of the dynamics of collective attention to political parties. *PLOS One*, 2015.

[110] Gunther Eysenbach. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4):e123, 2011.

[111] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *PloS one*, 9(10):e110184, 2014.

[112] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[113] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics, 2010.

[114] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[115] Marcel Flores and Aleksandar Kuzmanovic. Searching for spam: detecting fraudulent accounts via web search. In *Passive and Active Measurement*, pages 208–217. Springer, 2013.

[116] followerwonk. http://followerwonk.com/.

[117] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393*, 2018.

[118] Linton Freeman. The development of social network analysis. *A Study in the Sociology of Science*, 1, 2004.

[119] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[120] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Sampling online social networks: an experimental study of twitter. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 127–128. ACM, 2014.

[121] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying social networks at scale: macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 277–288. ACM, 2014.

[122] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203–1233, 2000.

[123] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok Choudhary. Towards online spam filtering in social networks. In *Symposium on Network and Distributed System Security (NDSS)*, 2012.

[124] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of ACM SIGCOMM conference on Internet measurement - IMC '10*, page 35, New York, New York, USA, 2010. ACM Press.

[125] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *CoRR*, abs/1206.5851, 2012.

[126] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'11*. The AAAI Press, 2011.

[127] Duncan Geere. It's not just you: 71 percent of tweets are ignored. https://bit.ly/2H9hZTr, 2010. [Online; accessed 2018-30-12].

[128] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. Understanding and combating link farming in the Twitter social network. In *Proceedings of the international conference on World Wide Web - WWW '12*, page 61, New York, New York, USA, apr 2012. ACM Press.

[129] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

[130] Zafar Gilani, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. Stweeler: A framework for twitter bot analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 37–38. International World Wide Web Conferences Steering Committee, 2016.

[131] CJ Hutto Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media- ICWSM'11)*, 2014.

[132] Github. Elections study. `https://github.com/antonak/elections_study/blob/0.1/study.ipynb`. [Online; Accessed on 3 February 2020].

[133] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[134] N Godbole and M Srinivasaiah. Large-scale sentiment analysis for news and blogs. *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'07*, pages 219–222, 2007.

[135] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1105–1108, New York, NY, USA, 2011. ACM.

[136] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on Twitter networks: validation of Dunbar's number. *PloS one*, 6(8):e22656, jan 2011.

[137] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[138] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security - CCS '10*, page 27, New York, New York, USA, October 2010. ACM Press.

[139] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the ACM Conference on Computer and Communications Security - CCS'10*, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM.

[140] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, pages 5228–35, apr 2004.

[141] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

[142] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[143] Robert A Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[144] Del Harvey. Trust And Safety. https://bit.ly/2LWt3Cl, 2010. [Online; Accessed: 2018-12-30].

[145] Taher H. Haveliwala and Taher H. Topic-sensitive PageRank. In *Proceedings of the international conference on World Wide Web - WWW '02*, page 517, New York, New York, USA, 2002. ACM Press.

[146] Yulan He, Hassan Saif, Zhongyu Wei, and Kam-fai Wong. Quantising opinions for political tweets analysis. In *International Conference on Language Resources and Evaluation - LREC'12*. Citeseer, 2012.

[147] Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina, Victor Martinez-Hernandez, Victor Sanchez, and Héctor Perez-Meana. A web scraping methodology for bypassing twitter api restrictions. *arXiv preprint arXiv:1803.09875*, 2018.

[148] J E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–72, nov 2005.

[149] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in Twitter. In *Proceedings of the international conference companion on World wide web - WWW '11*, page 57, New York, New York, USA, 2011. ACM Press.

[150] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[151] Sounman Hong and Daniel Nadler. Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates

and its impact on candidate salience. *Government Information Quarterly*, 29(4):455–461, 2012.

[152] Daniel Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 01/2010 2010.

[153] Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. Conversational Tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[154] InternetLiveStats.com. Twitter Usage Statistics - Internet Live Stats. www.internetlivestats.com/twitter-statistics/, 2018. [Online; Accessed: 2018-12-30].

[155] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the WebKDD and - SNA-KDD workshop on Web mining and social network analysis WebKDD/SNA-KDD'07*, pages 56–65. ACM, 2007.

[156] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160. Association for Computational Linguistics, 2011.

[157] Samuel Johnson, Joaquín J Torres, J Marro, and Miguel A Munoz. Entropic origin of disassortativity in complex networks. *Physical review letters*, 104(10):108702, 2010.

[158] Aditya Joshi, Prayas Jain, Pushpak Bhattacharyya, and Mark Carman. Who would have thought of that!': A hierarchical topic model for extraction of sarcasm-prevalent topics and sarcasm detection. *arXiv preprint arXiv:1611.04326*, 2016.

[159] Paul Judge. Barracuda Labs 2010, Annual security report. Technical report, Barracuda Networks Inc., 2010.

[160] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage. Spamalytics: an empirical analysis of spam marketing conversion. In *CCS '08: Proceedings of the 15th ACM conference on Computer and communications security*, pages 3–14, New York, NY, USA, 2008. ACM.

[161] Alex Kantrowitz. How twitter made the tech world's most unlikely comeback. https://bit.ly/2M0sOpy. [Online; Accessed: 2018-10-21].

[162] Amir Karami, Alicia A Dahl, Gabrielle Turner-McGrievy, Hadi Kharrazi, and George Shaw. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38(1):1–6, 2018.

[163] E Kim, S Gilbert, M Edwards, and E Graeff. Detecting sadness in 140 characters. *Webecology Project*, 2009.

[164] JM Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, aug 2000.

[165] Kaj-Kolja Kleineberg and Marián Boguñá. Evolution of the digital society reveals balance between viral and mass media influence. *Phys. Rev. X*, 4:031046, Sep 2014.

[166] Olga Kolchyna, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015.

[167] Bence Kollanyi, Philip N Howard, and Samuel C Woolley. Bots and automation over twitter during the first us presidential debate. Technical report, COMPROP Data Memo, 2016.

[168] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the International AAAI Conference on Weblogs and Social Media- ICWSM'11*, 11(538-541):164, 2011.

[169] Brian Krebs. Twitter bots drown out anti-kremlin tweets. https://krebsonsecurity.com/tag/maxim-goncharov/. [Online; Accessed: 2018-12-30].

[170] Christian Kreibich, Chris Kanich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, Vern Paxson, and Stefan Savage. On the spam campaign trail. *LEET*, 8:1–9, 2008.

[171] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.

[172] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 611, New York, New York, USA, aug 2006. ACM Press.

[173] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 905, New York, New York, USA, 2007. ACM Press.

[174] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2335–2338. ACM, 2012.

[175] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the international conference on World wide web - WWW '10*, page 591, New York, New York, USA, April 2010. ACM Press.

[176] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[177] Peter Laflin, Alexander V Mantzaris, Fiona Ainley, Amanda Otley, Peter Grindrod, and Desmond J Higham. Discovering and validating influence in a dynamic online social network. *Social Network Analysis and Mining - SNAM'13*, 3(4):1311–1323, 2013.

[178] Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 993–1003, 2013.

[179] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, page 435, New York, New York, USA, 2010. ACM Press.

[180] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'11*, 2011.

[181] Emil Leong. New Ways to Control Your Experience on Twitter. https://bit.ly/2b2dtRD, 2016. [Online; Accessed: 2018-12-30].

[182] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'10*, pages 90–97, 2010.

[183] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, page 462, New York, New York, USA, August 2008. ACM Press.

[184] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 631, New York, New York, USA, August 2006. ACM Press.

[185] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '05*, page 177, New York, New York, USA, August 2005. ACM Press.

[186] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. In *TKDD*, volume 1, page 2. ACM, 2007.

[187] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the international conference on World Wide Web - WWW '08*, page 695, New York, New York, USA, April 2008. ACM Press.

[188] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM, 2012.

[189] C. Liebrecht, F. Kunneman, and A. van den Bosch. The perfect solution for detecting sarcasm in tweets or not. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis - SSA'13*. Association for Computational Linguistics, Jun 2013.

[190] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Aaai*, volume 12, pages 22–26, 2012.

[191] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen Ebrahimi Moghaddam, and Lyle H Ungar. Analyzing personality through social media profile picture choice. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'16*, pages 211–220, 2016.

[192] Avishay Livne, Matthews P. Simmons, W. Abraham Gong, Eytan Adar, and Lada A. Adamic. The party is over here: structure and content in the 2010 election. Association for the Advancement of Artificial Intelligence, 2011.

[193] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. Estimating collective judgement of rumours in social media. *CoRR*, abs/1506.00468, 2015.

[194] Kamesh Madduri, David Ediger, Karl Jiang, David A. Bader, and Daniel Chavarria-Miranda. A faster parallel algorithm and efficient multithreaded implementations for evaluating betweenness centrality on massive datasets. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pages 1–8. IEEE, may 2009.

[195] Warih Maharani, Alfian Akbar Gozali, et al. Degree centrality and eigenvector centrality in twitter. In *International Conference on Telecommunication Systems Services and Applications - TSSA'14*, pages 1–5. IEEE, 2014.

[196] Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. Did you take the pill?-detecting personal intake of medicine from twitter. *arXiv preprint arXiv:1808.02082*, 2018.

[197] Carolanne Mangles. Search Engine Statistics 2018. `https://bit.ly/2Bwhqva`, 2018. [Online; Accessed: 2018-12-30].

[198] Evangelos Markatos, Davide Balzarotti, Magnus Almgren, Elias Athanasopoulos, Herbert Bos, Lorenzo Cavallaro, Sotiris Ioannidis, Martina Lindorfer, Federico Maggi, Zlatogor Minchev, et al. *The Red Book*. SysSec Consortium, 2013.

[199] marketingcharts. Social Networking Eats Up 3+ Hours Per Day For The Average American User. `https://bit.ly/1mmPPhB`, 2013. [Online; Accessed: 2018-12-30].

[200] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2):431–441, 1963.

[201] Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, L Alfonso Urena-López, and A Rturo Montejo-Ráez. Sentiment analysis in twitter. *Natural Language Engineering*, 20(1):1–28, 2014.

[202] Juan Martinez-Romo and Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000, 2013.

[203] Juan Martinez-Romo and Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications - Expert Syst. Appl.'13*, 40(8):2992–3000, June 2013.

[204] Christopher Matthews. How does one fake tweet cause a stock market crash? `https://bit.ly/2FkPjEE`, 2013. [Times.com, Accessed: 2018-12-30].

[205] DG Maynard and Mark A Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *International Conference on Language Resources and Evaluation - LREC'14*. ELRA, 2014.

[206] Diana Maynard and Mark A Greenwood. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. *Lrec*, pages 4238–4243, 2014.

[207] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[208] Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M. Voelker, Stefan Savage, and Kirill Levchenko. PharmaLeaks: understanding the business of online pharmaceutical affiliate programs, 2012.

[209] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. On building a reusable twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1113–1114. ACM, 2012.

[210] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

[211] Brendan Meeder, Brian Karrer, Amin Sayedi, R. Ravi, Christian Borgs, and Jennifer Chayes. We know who you followed last summer: Inferring social link creation times in twitter. In *Proceedings of the International Conference on World Wide Web - WWW'11*, WWW '11, pages 517–526, New York, NY, USA, 2011. ACM.

[212] Brendan Meeder, Brian Karrer, Amin Sayedi, R Ravi, Christian Borgs, and Jennifer Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on World wide web*, pages 517–526. ACM, 2011.

[213] Yelena Mejova, Padmini Srinivasan, and Bob Boynton. Gop primary season on twitter : popular political sentiment in social media. In *Proceedings of the ACM international conference on Web search and data mining - WSDM'13*, pages 517–526. ACM, 2013.

[214] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 71–79, New York, New York, USA, 2010. ACM Press.

[215] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12*, pages 379–387. ACM, 2012.

[216] Vassili van der Mersch. Twitter's 10 year struggle with developer relations. `https://bit.ly/2TAG1YR`. Accessed: 2018-10-28.

[217] Johnnatan Messias, Lucas Schmidt, Ricardo Oliveira, and Fabrício Benevenuto. You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18(7), 2013.

[218] Anjali Midha. Study: Exposure to brand tweets drives consumers to take action - both on and off twitter. `https://bit.ly/2CgY6UV`, 2014. [Online; accessed 2018-30-12].

[219] George Mikros and Kostas Perifanos. Authorship attribution in greek tweets using author's multilevel n-gram profiles. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'13*, 2013.

[220] Stanley Milgram. The small world problem. In *Psychology today*, volume 2, pages 60–67. New York, 1967.

[221] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of ACM SIGCOMM conference on Internet measurement - IMC '07*, page 29, New York, New York, USA, October 2007. ACM Press.

[222] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417, 2013.

[223] Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499, 2015.

[224] AJ Morales, J Borondo, JC Losada, and RM Benito. Efficiency of human activity on information spreading on twitter. In *Elsevier - Social Networks*, volume 39, pages 1–2011. Elsevier, 2014.

[225] AJ Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[226] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

[227] Margi Murphy. Twitter to remove ‘like’ tool in a bid to improve the quality of debate. https://bit.ly/2yExMmK. [Online; Accessed: 2018-11-15].

[228] Paul Mutton and Jennifer Golbeck. Visualization of semantic metadata and ontologies. In *Proceedings of International Conference on Information Visualization - InfoVis’03*, pages 300–305. IEEE, 2003.

[229] Louise Myers. What Happens in a Twitter Minute? Infographic. https://louisem.com/6267/twitter-minute-infographic, 2014. [Online; Accessed: 2018-12-30].

[230] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: The structure of the twitter follow graph. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 493–498. International World Wide Web Conferences Steering Committee, 2014.

[231] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[232] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the international workshop on semantic evaluation - SemEval’16*, pages 1–18, 2016.

[233] Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12–14, 2012.

[234] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference*, page 8. ACM, 2011.

[235] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[236] Mark EJ Newman. Power laws, pareto distributions and zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.

[237] Todd P Newman. Tracking the release of ipcc ar5 on twitter: Users, comments, and sources following the release of the working group i summary for policymakers. *Public Understanding of Science*, 26(7):815–825, 2017.

[238] World News. Top 10 social media disasters of 2017. https://www.theweek.co.uk/90348/top-10-social-media-disasters-of-2017, 2017.

[239] Ryosuke Nishi, Taro Takaguchi, Keigo Oka, Takanori Maehara, Masashi Toyoda, Kenichi Kawarabayashi, and Naoki Masuda. Reply trees in twitter : data analysis and branching process models. *Social Network Analysis and Mining - SNAM'16*, 6(1):26, 2016.

[240] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[241] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'10*, 2010.

[242] John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer, and Sibel Adalii. Credibility in context: An analysis of feature distributions in twitter. In *SocialCom/PASSAT*, pages 293–301, 2012.

[243] Ozer Ozdikis, Pinar Senkul, and Halit Oguztuzun. Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*, 2012.

[244] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[245] Shay Palachy. A list of twitter datasets and related resources. https://bit.ly/2H5P8zu, 2018. [Online; Accessed 2018-12-30].

[246] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008.

[247] Nikolaos Papatheodorou, Pepi Stavropoulou, Dimitrios Tsonos, Georgios Kouroupetroglou, Dimitris Spiliotopoulos, and Charalambos Papageorgiou. *On the Identification and Annotation of Emotional Properties of Verbs*, chapter On the Move to Meaningful Internet Systems: OTM 2013, Springer, p. 588-597, pages 588–597. Springer, 2013.

[248] Peter F Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z Pan, Ian Horrocks, and Birte Glimm, editors. *Making Sense of Twitter*, pages 470–485. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[249] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[250] Alberto Pepe and Johan Bollen. Between conjecture and memento: Shaping a collective emotional perception of the future. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, pages 111–116, 2008.

[251] N Perlroth. Fake twitter followers become multimillion-dollar business. *The New York Times*, 2013. [Online; Accessed: 2018-12-30].

[252] Saša Petrović, Miles Osborne, and Victor Lavrenko. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[253] René Pfitzner, Antonios Garas, and Frank Schweitzer. Emotional divergence influences information spreading in twitter. *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM'12)*, 12:2–5, 2012.

[254] Christopher Potts. Sentiment Symposium Tutorial: Lexicons. https://bit.ly/2smM9Zo, 2011. [Online; Accessed 2018-12-30].

[255] Polyvios Pratikakis. twawler: A lightweight twitter crawler. *arXiv preprint arXiv:1804.07748*, 2018.

[256] Daniel Preotiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. Studying user income through language, behaviour and affect in social media. *PLOS One*, September 2015.

[257] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE, 2011.

[258] Marufur Rahman and Rezaul Karim. Comparative study of different methods of social network analysis and visualization. In *International Conference on Networking Systems and Security - NSysS'16*, pages 1–7. IEEE, 2016.

[259] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. Detecting and Tracking Political Abuse in Social Media. *Proceedings of the International AAAI Conference on Weblogs and Social Media - ICWSM11)*, 2011.

[260] JL Reiss. Statistical methods for rates and proportions. *Second Edit. John Wiley and Sons, New York*, pages 212–225, 1981.

[261] Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V Zicari. Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems - KBS'14*, 69:24–33, 2014.

[262] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. pages 704–714. ACL, 2013.

[263] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[264] Giuseppe Rizzo and Raphaël Troncy. Nerd: A framework for evaluating named entity recognition tools in the web of data. In *International Semantic Web Conference - ISWC'11*, pages 1–4, 2011.

[265] Glívia Angélica Rodrigues Barbosa, Ismael S. Silva, Mohammed Zaki, Wagner Meira, Jr., Raquel O. Prates, and Adriano Veloso. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 2621–2626, New York, NY, USA, 2012. ACM.

[266] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *Proceedings of the international conference companion on World wide web - WWW '11*, page 113, New York, New York, USA, 2011. ACM Press.

[267] Aliza Rosen and Ikuhiro Ihara. Giving you more characters to express yourself. `https://bit.ly/2fQ2b7W`, 2018. [Twitter.com, Accessed: 2018-12-30].

[268] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[269] Yoel Roth and Del Harvey. How twitter is fighting spam and malicious automation. https://bit.ly/2N4OumE. [Online; Accessed: 2018-10-20].

[270] Eldar Sadikov and Maria Montserrat Medina Martinez. Information propagation on twitter. *CS322 project report*, 2009.

[271] Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. In *2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on theWorld Wide Web (WWW'12)*, pages 2–9. CEUR Workshop Proceedings (CEUR-WS.org), 2012.

[272] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *International Semantic Web Conference - ISWC'12*, pages 508–524. Springer, 2012.

[273] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the workshop on semantic analysis in social media*, pages 53–60, 2012.

[274] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM, 2015.

[275] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787, 2018.

[276] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, and Chengshan Zhang. An empirical analysis of phishing blacklists. In *Proceedings of Conference on Email and Anti-Spam - CEAS'09*, 2009.

[277] Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, and Jacob Spoelstra. Predicting us primary elections with twitter. https://stanford.io/2shORiz, 2012. Accessed: 2018-12-30.

[278] Xin Shuai, Alberto Pepe, and Johan Bollen. How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one*, 7(11):e47523, 2012.

[279] Carson Sievert and Kenneth E. Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014.

[280] Marko Skoric, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim, and Jing Jiang. Tweets and votes: A study of the 2011 singapore general election. In *International Conference on System Science - HICSS'2012*, pages 2583–2591. IEEE, 2012.

[281] Aaron Smith and Monica Anderson. Social Media Use in 2018. https://pewrsr.ch/2FDfiFd, 2018. [Online; Accessed: 2018-12-30].

[282] Bryor Snefjella, Daniel Schmidtke, and Victor Kuperman. National character stereotypes mirror language use: A study of canadian and american tweets. *PLOS ONE*, 13(11):1–37, 11 2018.

[283] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[284] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 53–63, 2011.

[285] Vasumathi Sridharan, Vaibhav Shankar, and Minaxi Gupta. Twitter games: How successful spammers pick targets. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC '12, pages 389–398, New York, NY, USA, 2012. ACM.

[286] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences - PNAS'18*, page 201803470, 2018.

[287] Brett Stone-Gross, Ryan Abman, Richard A. Kemmerer, Christopher Kruegel, Douglas G. Steigerwald, and Giovanni Vigna. The Underground Economy of Fake Antivirus Software. In *Economics of Information Security and Privacy - WEIS'13*, pages 55–78. Springer New York, New York, NY, 2013.

[288] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[289] Gianluca Stringhini, Gang Wang, Manuel Egele, Christopher Kruegel, Giovanni Vigna, Haitao Zheng, and Ben Y Zhao. Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 163–176. ACM, 2013.

[290] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268, 2001.

[291] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, et al. The darpa twitter bot challenge. *arXiv preprint arXiv:1601.05140*, 2016.

[292] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *IEEE International Conference on Social Computing - SocialCom'10*, pages 177–184. IEEE, 2010.

[293] Partha Pratim Talukdar and Koby Crammer. New Regularized Algorithms for Transductive Learning. In *Machine Learning and Knowledge Discovery in Databases - ECML PKDD'09*, pages 442–457. Springer Berlin Heidelberg, 2009.

[294] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the international workshop on semantic evaluation - SemEval'14*, pages 208–212, 2014.

[295] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL'14*, volume 1, pages 1555–1565, 2014.

[296] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the ACM international conference on Web search and data mining - WSDM'11*, pages 35–44. ACM, 2011.

[297] Mike Thelwall and Kevan Buckley. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. In *Journal of the American Society for Information Science and Technology - JASIST'13*, volume 64, pages 1608–1617, 2013.

[298] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173, January 2012.

[299] Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5):e64841, 2013.

[300] Kurt Thomas, Chris Grier, and Vern Paxson. Adapting social spam infrastructure for political censorship. In *USENIX Security Symposium - USENIX'12*, 2012.

[301] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of ACM SIGCOMM conference on Internet measurement - IMC '11*, pages 243–258, New York, NY, USA, 2011. ACM.

[302] Kurt Thomas, Frank Li, Chris Grier, and Vern Paxson. Consequences of Connectivity. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, pages 489–500, New York, New York, USA, nov 2014. ACM Press.

[303] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *USENIX Security Symposium - USENIX'13*, 2013.

[304] James Titcomb. Twitter makes first profit in 12-year history. https://bit.ly/2RD1MtD. [telegraph.co.uk, Accessed: 2018-11-15].

[305] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *JSTOR - Sociometry*, pages 425–443. JSTOR, 1969.

[306] Erik Tromp and Mykola Pechenizkiy. Senticorr: Multilingual sentiment analysis of personal correspondence. In *IEEE International Conference on Data Mining Workshops - ICDMW'11*, pages 1247–1250. IEEE, 2011.

[307] Adam Tsakalidis, Symeon Papadopoulos, and Ioannis Kompatsiaris. An ensemble model for cross-domain polarity classification on twitter. In *International Conference on Web Information Systems Engineering - WISE'14*, pages 168–177, 2014.

[308] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social science computer review - SSC'11*, 29(4):402–418, 2011.

[309] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter : What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Weblogs and Social Media- ICWSM'10*, 10(1):178–185, 2010.

[310] Twitter. How to use Twitter lists. https://help.twitter.com/en/using-twitter/twitter-lists. [Online; Accessed on 18 June 2019].

[311] Twitter Help Center. The twitter rules. https://bit.ly/2j9xU9n. [Online; Accessed: 2018-12-30].

[312] Twitter Inc. Shutting down spammers. https://bit.ly/2VEEZx1. Twitter.com, Accessed: 2018-12-30.

[313] Twitter official API documentation. Standard api rate limits per window. https://bit.ly/2REDPCl. [Online; Accessed: 2018-11-15].

[314] Twitter official blog. Delivering a consistent twitter experience. https://bit.ly/2C8KX00. [Online; Accessed: 2018-11-15].

[315] Twitter Official Blog. Continuing our commitment to health. https://bit.ly/2tocAOi, 2018. [Online; Accessed 2018-12-30].

[316] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[317] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science - MIT Initiative on the Digital Economy - IDE'18*, 359(6380):1146–1151, 2018.

[318] Alex Hai Wang. Don't follow me - spam detection in twitter. In *International Conference on Security and Cryptography - SECRYPT'10*, pages 142–151, 2010.

[319] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL System Demonstrations*, pages 115–120. Association for Computational Linguistics -ACL'12, 2012.

[320] Tianyi Wang, Yang Chen, Zengbin Zhang, Peng Sun, Beixing Deng, and Xing Li. Unbiased sampling in directed social graph. *ACM SIGCOMM Computer Communication Review - SIGCOMM'11*, 41(4):401–402, 2011.

[321] Yu Wang, Yang Feng, Zhe Hong, Ryan Berger, and Jiebo Luo. How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International Conference on Social Informatics - Socinfo'19*, pages 440–456. Springer, 2017.

[322] Yuan Wang, Jie Liu, Jishi Qu, Yalou Huang, Jimeng Chen, and Xia Feng. Hashtag graph based topic model for tweet mining. In *IEEE International Conference on Data Mining - ICDM'14*, pages 1025–1030. IEEE, 2014.

[323] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop - NAACL'16*, pages 88–93, 2016.

[324] Benjamin Waugh, Maldini Abdipanah, Omid Hashemi, Shaquille Abdul Rahman, and David M Cook. The influence and deception of twitter: the authenticity of the narrative and slacktivism in the australian electoral process. *ECCWS2014-Proceedings of the 13th European Conference on Cyber warefare and security*, 2013.

[325] Benjamin Waugh, Maldini Abdipanah, Omid Hashemi, Shaquille Rahman, and David Cook. The Influence and Deception of Twitter: The Authenticity of the Narrative and Slacktivism in the Australian Electoral Process, 2013.

[326] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining - ASONAM'13*, pages 290–297. ACM, 2013.

[327] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank. In *Proceedings of the ACM international conference on Web search and data mining - WSDM '10*, page 261, New York, New York, USA, February 2010. ACM Press.

[328] Ryan Wesslen, Sagar Nandu, Omar Eltayeby, Tiffany Gallicano, Sara Levens, Min Jiang, and Samira Shaikh. Bumper stickers on the twitter highway: Analyzing the speed and substance of profile changes, Jun 2018.

[329] Leo Widrich. How twitter evolved from 2006 to 2011. "https://blog.bufferapp.com/how-twitter-evolved-from-2006-to-2011", 2011. [Online; accessed 10-October-2015].

[330] wikipedia. 2015 greek bailout referendum. https://en.wikipedia.org/wiki/2015_Greek_bailout_referendum. [Online; Accessed on 11/10/2019].

[331] Wikipedia. Aol. https://en.wikipedia.org/wiki/AIM_(software). [Online; Accessed on 18 June 2019].

[332] wikipedia. Dnsbl. https://en.wikipedia.org/wiki/DNSBL. [Online; Accessed on 11/10/2019].

[333] Wikipedia. facebook. https://en.wikipedia.org/wiki/facebook. [Online; Accessed on 18 June 2019].

[334] Wikipedia. instagram. https://en.wikipedia.org/wiki/instagram. [Online; Accessed on 18 June 2019].

[335] Wikipedia. List of social networking websites. https://bit.ly/1my8Jr1. [Online; Wikipedia.org, Accessed: 2018-10-27].

[336] Wikipedia. myspace. https://en.wikipedia.org/wiki/myspace. [Online; Accessed on 18 June 2019].

[337] Wikipedia. Occupy wall street. https://en.wikipedia.org/wiki/Occupy_Wall_Street. [Online; Accessed on 1 August 2019].

[338] Wikipedia. Plato. https://en.wikipedia.org/wiki/PLATO_(computer_system). [Online; Accessed on 18 June 2019].

[339] Wikipedia. Random walk. https://en.wikipedia.org/wiki/Random_walk. [Online; Accessed on 1 August 2019].

[340] Wikipedia. Social media and the arab spring. https://en.wikipedia.org/wiki/Social_media_and_the_Arab_Spring. [Online; Accessed on 1 August 2019].

[341] Wikipedia. spam. https://en.wikipedia.org/wiki/Spamming.

[342] Wikipedia. Talkomatic. https://en.wikipedia.org/wiki/Talkomatic. [Online; Accessed on 18 June 2019].

[343] Wikipedia. Timeline of social media. https://en.wikipedia.org/wiki/Timeline_of_social_media. [Online; Accessed on 18 June 2019].

[344] Wikipedia. Windows live messenger. https://en.wikipedia.org/wiki/Windows_Live_Messenger. [Online; Accessed on 18 June 2019].

[345] Wikipedia. Timeline of twitter. "https://en.wikipedia.org/wiki/Timeline_of_Twitter", 2004. [Online; accessed 10-October-2015].

[346] Wikipedia. Psychometrics. https://en.wikipedia.org/wiki/Psychometrics, 2018. [Wikipedia.org, Accessed: 2018-12-30].

[347] Chester Wisniewski. Twitter hack demonstrates the power of weak passwords. https://bit.ly/2sgQsFi, 2010. [Online;Accessed: 2018-12-30].

[348] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on Twitter . In *Proceedings of the international conference on World wide web - WWW '11*, page 705, New York, New York, USA, 2011. ACM Press.

[349] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the ACM international conference on Web search and data mining - WSDM'11*, pages 177–186. ACM, 2011.

[350] Shaozhi Ye and S Felix Wu. Measuring message propagation and social influence on twitter. com. In *International Conference on Social Informatics - Socinfo'10*, pages 216–231. Springer, 2010.

[351] Shaozhi Ye and Shyhtsun Felix Wu. Measuring message propagation and social influence on twitter.com. *International Conference on Social Informatics - Socinfo'10*, 10:216–231, 2010.

[352] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. SybilLimit: A Near-Optimal Social Network Defense Against Sybil Attacks. *IEEE/ACM Transactions on Networking - TON'10*, 18(3):885–898, jun 2010.

[353] Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. In *ACM SIGCOMM Computer Communication Review - SIGCOMM'06*, volume 36, pages 267–278. ACM, 2006.

[354] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

[355] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, pages 1528–1531. ACM, 2012.

[356] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European conference on information retrieval - ECIR'11*, pages 338–349. Springer, 2011.

[357] Bin Zou, Vasileios Lampos, Russell Gorton, and Ingemar J Cox. On infectious intestinal disease surveillance using social media content. In *Proceedings of the International Conference on Digital Health Conference- ICDHT'16*, pages 157–161. ACM, 2016.

[358] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*, 2016.