Doctoral Thesis

Potential Target Audience of Misinformation on Social Media: Credulous Users

PhD Program in Computer Science: XXXII cycle

Supervisor:
Prof. Rocco De Nicola
rocco.denicola@imtlucca.it

Doctoral Candidate:

Alessandro Balestrucci
alessandro.balestrucci@gssi.it

Co-supervisors:
Dr. Marinella Petrocchi
marinella.petrocchi@iit.cnr.it

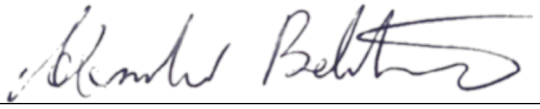
Dr. Catia Trubiani
catia.trubiani@gssi.it

GSSI Gran Sasso Science Institute
Viale Francesco Crispi, 7 - 67100 L'Aquila - Italy

Declaration of Authorship

I, Alessandro Balestrucci, declare that this thesis titled, 'Potential target audience of misinformation on Social Media: the Credulous users' and the work presented in it are my own under the guidance of my supervisors Prof. De Nicola Rocco, Dr. Petrocchi Marinella and Dr. Trubiani Catia. Specifically, Chapter 3 is based on [11], co-authored with Prof. De Nicola, Dr. Trubiani Catia and Dr. Inverso Omar. Chapter 4 is based on [12], co-authored with Prof. De Nicola Rocco, Dr. Petrocchi Marinella and Dr. Trubiani Catia. Chapter 5 is based on [8]. Chapter 6 is based on [9], co-authored with Prof. De Nicola Rocco, Dr. Petrocchi Marinella and Dr. Trubiani Catia. Chapter 7 is based on [10], co-authored with Prof. De Nicola Rocco.
Stated that, I further confirm the following:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 10/12/2020

A mia madre, mio padre e i miei fratelli Giuseppe e Luciano...
e a coloro che ormai da tempo mi guardano da lassù... i miei nonni

To my mom, my dad and my brothers Giuseppe and Luciano...
and to those who never ceased looking upon me... my grandparents

"Do not judge me by my successes, judge me by how many times I fell down and got back up again."

Nelson Mandela

"Strength does not come from physical capacity. It comes from an indomitable will."

Gandhi

Acknowledgements

It is with a certain emotion that I am going to write this paragraph, in which it is my pleasure to remember and thank all those who, more or less significantly and in different ways, have contributed to this work. During my years of doctoral studies, I have been lucky to meet and work with wonderful and unforgettable people who made this experience unique.

First of all, I want to thank my supervisor, Prof. Rocco De Nicola (IMT Lucca). I am especially grateful to him for his perseverance, patience and for sharing with me his knowledge and experience, as well as for his suggestions and advice. It has been a privilege to have been his student and to work with him.
I thank my co-supervisors, Dr. Marinella Petrocchi and Dr. Catia Trubiani, for their collaboration, the uncountable discussions, their willingness to assist me, their suggestions and for all the help I received especially in moments of difficulty. Thank you.

I thank all computer scientists, be they post-docs, researchers or professors. I would especially like to thank Prof. Michele Flammini, and Prof. Luca Aceto, their kindness, capacity to listen, sensitivity and competence have been an inspiration to everybody.
I thank the IT staff of GSSI, the IT division of LNGS (ULITE) and all those who have provided me with the IT resources that have contributed to the results reported in this document. Without them, it would have been almost impossible to carry out the experiments and analyses of this work.
Thanks and a big hug to my course (Computer Science) colleagues, for their friendship, constant support, advice and for all the good times spent together. Knowing you has made me richer as a human being.

I thank the Gran Sasso Science Institute and IMT Lucca, the two institutes that allowed me to start and finish my doctoral studies. I thank the European Union programme Horizon 2020 (grant agreement n. 830892, SPARTA) and the integrated activity project TOFFEe 'TOols for Fighting FakEs' (IMT Lucca) for all the support given to me.

I thank the Lovreglio family, my cousins in L'Aquila. Not many can say to have attended their doctorate in a city different than their own and have found relatives whom they could ask for help. I consider it the greatest fortune that has happened to me since the beginning of this experience. Thanks to them I missed home less.
The biggest thanks goes to my parents (Maria Sterpeta and Andrea) and both my brothers Giuseppe and Luciano, for having transfered to me the serenity that was necessary to go on, and for having always believed in me; especially when I was discouraged and I had lost my self-confidence. Many times I have doubted of myself and about success;

they were the ones who have motivated, encouraged, convinced me not to give up and to always go on. If today I was able to end up this journey, I owe it especially to them.

Ringraziamenti

E' con una certa commozione che mi accingo a scrivere questo paragrafo, nel quale è mio piacere ricordare e ringraziare tutti coloro che, più o meno significativamente e in modi diversi, hanno contribuito a questo lavoro. Durante questi anni di dottorato, ho avuto la fortuna di incontrare e lavorare con persone meravigliose e indimenticabili che hanno reso questa esperienza unica.

In primis voglio ringraziare il mio supervisor, Prof. Rocco De Nicola (IMT Lucca). Oltre che per i suoi suggerimenti e consigli, gli sono particolarmente grato per la sua costanza, pazienza e per aver condiviso con me la sua conoscenza ed esperienza. E' stato un privilegio essere stato suo studente e lavorare con lui.
Ringrazio i miei co-supervisors, Dr. Marinella Petrocchi e Dr. Catia Trubiani, per la loro collaborazione, le innumerevoli conversazioni, la loro volontà nell'assistermi, i loro suggerimenti e per tutto l'aiuto ricevuto specialmente nei momenti di difficoltà. Grazie.

Ringrazio tutti gli uomini e donne di computer science nelle loro rispettive cariche di post-doc, ricercatori e professori. In particolare, voglio ringraziare il Prof. Michele Flammini, e il Prof. Luca Aceto, la loro gentilezza, capacità di ascolto, sensibilità, professionalità e competenza è stata d'esempio per tutti.
Ringrazio lo staff IT del GSSI, la divisione IT di LNGS (ULITE) e tutti coloro che mi hanno fornito le risorse informatiche e di calcolo che hanno contribuito ai risultati riportati nel presente documento. Senza di loro sarebbe stata quasi impossibile la realizzazione degli esperimenti e delle analisi di questo lavoro.
Un ringraziamento e un grande abbraccio ai miei colleghi del corso, per la loro amicizia, per il loro costante sostegno, per i loro consigli e per tutti i bei momenti passati insieme. Conoscervi mi ha reso umanamente più ricco.

Ringrazio il Gran Sasso Science Institute e IMT Lucca, i due istituti che mi hanno permesso di iniziare e terminare i miei studi di dottorato. Ringrazio il programma dell'Unione Europea Horizon 2020 (convenzione di sovvenzione n. 830892, SPARTA) e il progetto di attività integrata TOFFEe 'TOols for Fighting FakEs' (IMT Lucca) per tutto il supporto concessomi.

Ringrazio la famiglia Lovreglio, i miei cugini de L'Aquila, non sono in molti a poter dire di aver frequentato il dottorato in una città diversa dalla propria e di aver trovato dei parenti a cui chiedere aiuto. La considero la più grande fortuna che mi sia capitata dall'inizio di questa esperienza. Grazie a loro ho sentito meno la mancanza di casa.

Il ringraziamento più grande va ai miei genitori (Maria Sterpeta e Andrea) ed entrambi i miei fratelli Giuseppe e Luciano per avermi trasmesso la serenità necessaria ad andare

avanti e per aver creduto sempre e comunque in me quando io stesso ero sfiduciato e avevo smarrito l'autostima. Più volte ho dubitato di me stesso e nella riuscita di questo percorso, loro sono stati coloro che mi hanno rimotivato, incoraggiato, convinto a non demordere e ad andare sempre avanti. Se oggi ho potuto terminare questo percorso lo devo specialmente a loro.

Abstract

The capability to reach a wider audience and the possibility to disseminate news faster are the main reasons for the growing importance of Online Social Media (OSM) whose usage has undoubtedly reshaped the way news are written, published and disseminated. However, due to the technical limits of online fact-checkers and to an uncontrolled content publishing, there is a high risk of being misinformed through fake news. Although automated accounts known as bots are considered the main promoters of mis-/dis- information diffusion, those who, with their actions, change the current events (e.g., welfare, economy, politics, etc.) are human users. Some categories of humans are more vulnerable to fake news than others, and performing mis-/dis- information activities targeting such categories would increase efficacy of such activities. Furthermore, recent studies have evidenced users' attitude not to fact-check the news they diffuse on OSM and thus the risk that they became themselves vectors of mis-/dis- information.

In this document, using Twitter as benchmark, we devote our attention to those human-operated accounts, named "credulous" users, which show a relatively high number of bots as *followees* (called *bot-followees*). We believe that these users are more vulnerable to manipulation (w.r.t. other human-operated accounts) and, although unknowingly, they can be involved in malicious activities such as diffusion of fake content. Specifically, we first design some heuristics by focusing on the aspects that best characterise the credulous users w.r.t. not credulous ones. Then, by applying Machine Learning (ML) techniques, we develop an approach based on binary classifiers able to automatically identify this kind of users and then use regression models to predict the percentage of humans' bot-followees (over their respective followees). Subsequently, we describe investigations conducted to ascertain the actual contribution of credulous users in the dissemination of potential malicious content and then, their involvements in fake news diffusion by analysing and comparing the fake news spread by credulous users w.r.t. not credulous one.

Our investigations and experiments, provide evidence of credulous users' involvement in spreading fake news thus supporting bots' actions on OSM.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Communication has always been among the basic needs of humanity. Since its first appearance in history (think of the prehistoric rock paintings) humans have wished to share information about their activities and interests with other fellow human beings.
The progress of technologies for information production and dissemination has played a pivotal role not only in improving communication effectiveness and dissemination but also in speeding up the processes of civilisation and modernisation. The invention of the modern press (Gutenberg 1455) has revolutionised and speeded up (compared to the amanuenses of that time) the information (hence, knowledge) production and dissemination fashion, access to which was considered a prerogative of clerical society (that had full control on it). All subsequent discoveries in (tele)communications, such as radio and television, have further expanded and strengthened not only the ways of informing and being informed but also the range of the audience. Then, with the birth of the World Wide Web, the *Digital Era* began and, with it, the unbounded spreading of information in the Digital World.

In the early days of the Internet, access to the network was limited to only those users who could afford purchasing a PC and an ISP (Internet Service Provider) subscription. In those years, these obstacles were far from trivial; but, over the years, the progress of technology and the creation of ad hoc infrastructures have allowed an increasing number of people to use the Net.
The growing community of internauts, combined with the possibility of communicating immediately regardless of users' geographical location, contributed to the appearance of social networking services to stimulate the aggregation of virtual communities, characterised by the possibility to relate users and allow them to exchange information. These platforms are known as *Online Social Media* (OSM).
It is under the definition of OSM that all those Web 2.0 [19] internet-based platforms fall,

given that they are offering users services able to facilitate the "[...] starting, sharing and exchanging of information and ideas in virtual communities and networks" [125]. Nowadays, the widespread use of mobile devices, combined with the ease and cheapness of being connected, are the most important factors increasing the usage, and hence the importance, of OSM as a means of communication [122]. The pervasiveness of OSM and their ease of use (facilitated by user-friendly applications) have lead to new ways for people to get informed. In fact, domestic users of popular social networking services, such as Twitter and Facebook, can keep up with the news effortlessly, while routinely checking out their own social channels of interest.
Several studies [40, 121–124] recently confirmed a growing trend (see Fig. 1.1), especially among young people [123] (see Fig. 1.2), in the use of OSM as the favourite information platform at the expense of traditional mass media, such as radio, newspaper and TV.

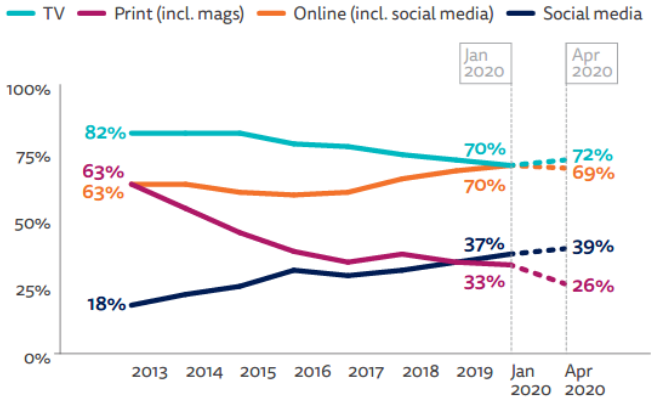


FIGURE 1.1: Mass Media usage trend in the last 7 years (image source [123]).

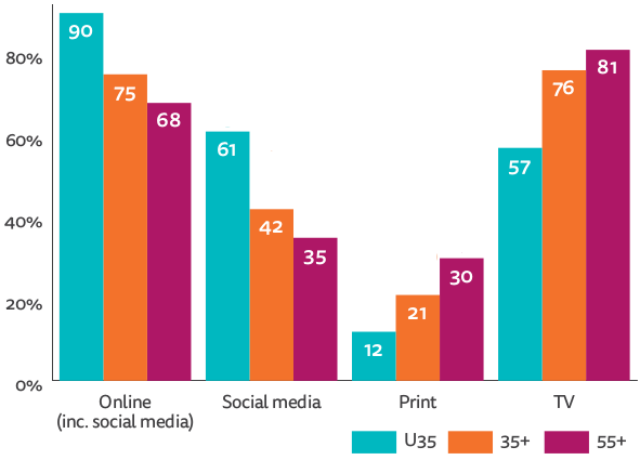


FIGURE 1.2: Comparison of media usage by users according to age (image source [123]).

Although these are very important advantages in a mass media, issues about content/news veracity, circulating on OSM, began to arise. The rapid proliferation of

user-generated content, the lack of tools capable of automating the fact-checking process, as well as the news sources reliability (as it happens in professional journalism in the most important newspapers or TV news), have allowed the publication and uncontrolled circulation of fake news. The use and abuse of these misleading and (often) harmful contents have led to the growth of mis- and dis- information phenomena.

Generally, the term "fake news" encompasses several types of totally or partially false news. Fake-news include, but are not limited to: satire, false connection, misleading content, false context, imposter content, manipulated content and fabricated content [173]. They are the building blocks to carry out campaigns of misinformation and disinformation. While the former uses fake news to arouse humour (e.g., satire, parody and stereotypes), therefore without any intention to harm, the latter (disinformation) aims to damage the image/reputation of a target. Alongside these two ways of misleading people, there is another one called malinformation. Sharing the same malevolent goals of disinformation, this kind of campaigns are orchestrated to damage the reputation or image of people, governments or organisations, through news from leaks (therefore potentially genuine but not officially confirmed) [174]. The differences between fake news, and the concepts of mis-/dis-/mal- information are explained in more detail in Section 2.2.1.

According to the 2019 report 'Weapons of mass distractions' [66], strategists of fake news can exploit (at least) three significant vulnerabilities of the online information ecosystem: i) the medium, ii) the message, and iii) the audience. As a matter of fact, the diffusion and the propagation of deliberately misleading information for harmful purposes is quite recurring in OSM, but the effectiveness of misinformation campaigns strongly depends on the ability to (i) attract people's interest by appropriate messages, and (ii) disseminated information.
The information diffusion on OSM is often supported by automated accounts totally (bots) or partially (cyborg) controlled by ad hoc software applications [71, 166]. In some cases, bots have been programmed for benevolent purposes, like: calling volunteers in case of emergencies [6, 144] or spreading academic events such as conferences and/or papers [75, 105], but these are only exceptions. Unfortunately, the dominant and worrisome use of these entities is far from being benign. Skillfully designed to mimic human behaviour online, such automated accounts interact (social bots [61]), under fictive identity, with genuine (in terms of being human) users and share/produce contents of doubtful credibility. Recent work [147, 183] demonstrates that bots are particularly active in spreading low credibility content and amplifying their significance. Typically operating in well-organised groups (called *botnets*), through the dissemination of misleading content, bots aim to pursue malicious purposes, e.g., to encourage hate speeches,

misconception, discontent and, more in general, to induce a bias within the public opinion [61, 88, 159, 183]. In fact, whatever the strategy adopted for spreading false news, this is only effective in presence of an audience willing to believe them.

In the survey 'A Report on the Spread of Fake News'[1,2] commissioned on 2017 by *Zignal Labs* (i.e., SaaS-based media intelligence software service company[3]) and conducted on 2,000 respondents, it is reported that:

1. 86% of respondents do not always fact check the articles they read via social media;
2. 61% of respondents are likely to comment on, like, or share content published by a friend;
3. 27% of respondents admittedly do not fact-check the articles they themselves share.

Moreover, from the findings in [89], where models for influence propagation in OSM have been studied by means of graphs, a strong correlation emerged between the target nodes (to influence) and the role of their neighbours (social contacts). Taking the above into consideration, it would not be preposterous to suspect that, depending on the activities of their social contacts, these users may well end up contributing actively, although unknowingly, to mis-/dis- information spreading; supporting, in such a way, bots' malicious activities on OSM. Moreover, let us consider the case where the human users' social contacts are mainly constituted by malicious bots; undoubtedly, this is a very worrisome scenario, especially if mis-/dis- information is performed in a targeted fashion by focusing such activities to an audience that, potentially, is easier to influence/deceive than another one (*targeted disinformation*).
Due to the impact that targeted disinformation can have on people, several governments began to consider it as a national security affair. In 2019 a team, working for USA Department of Homeland Security (DHS), provided the report 'Combatting Targeted Disinformation Campaigns'[4] where this phenomenon has been deeply investigated and formalised in a framework, known as *Disinformation Kill Chain*, which outlines the seven basic steps by which these campaigns are carried out (see Fig. 1.3).
Once a threat actor (e.g., botnet's masters) defines the goal of its disinformation campaign, the framework requires she/he performs the following steps resumed below:

- *Reconnaissance*: individuation of target audience, the medium (e.g., which OSM platform to use) and the arguments to exploit;
- *Build*: design and implement the infrastructure to use in the campaign (e.g., bots);

[1] A Report on the Spread of Fake News (survey): https://tinyurl.com/ybk3j55y
[2] 9 out of 10 Americans don't fact-check news they read on OSM: https://tinyurl.com/s67sq96
[3] Zignal Labs: https://zignallabs.com/
[4] Combatting Targeted Disinformation Campaigns: https://tinyurl.com/ybr4ntw2

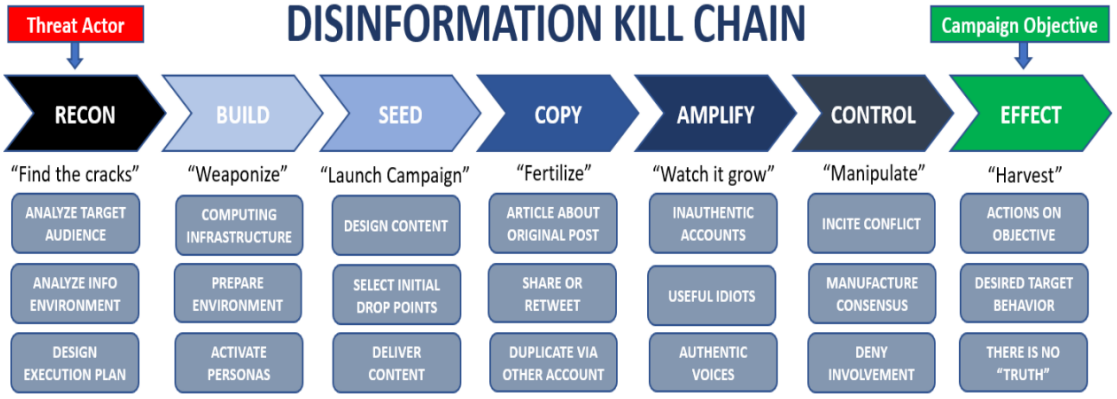


FIGURE 1.3: Disinformation Kill Chain (image source [4]).

- *Seed*: creation of fake/deceptive content and initial spreading (seeding) on OSM;
- *Copy*: production of content which refers to the original story, hence acting as an "information laundering", and making them to appear as authentic distribution;
- *Amplify*: making sure that the story ends up in the communication channels of the targeted audience. Bots and inauthentic accounts can help provide momentum, aiming to stimulate dissemination by other witting and *unwitting* agents. Successful amplification is achieved in case other unwitting agents and especially the target audience contribute to diffusing fake news to their peers;
- *Control*: pilot and manipulate the reactions of the targeted audience through interventions in conversations and debates to stimulate conflicts or obtain consensus;
- *Effect*: target audience begins to behave in line with the threat actor's objectives.

The action of audience targeting is quite recurring in *psychological warfare* (i.e., a branch of *information warfare* [99, 100]), also referred as *psychological operations*. This term is referred "to denote any action which is practiced mainly by psychological methods with the aim of evoking a planned psychological reaction in other people" [163]. Targeted disinformation is usually employed against ordinary people to cause some effect in their country's governments or also to 'stimulate' the population of other countries by means of technology and media [171][5,6,7].

The Oxford Internet Institute, in its study on the Global Inventory of Organised Social Media Manipulation (conducted in 2019), reports an increase of 150%in the number of countries using organised social media manipulation campaigns over the last two years: bot accounts are being used in 50 of the 70 investigated countries [22]; governments are not always the victims but, in some cases, also the (*threat*) actors[8] (in red in Fig. 1.3).

[5]What We Know—and Don't Know—About Facebook, Trump, and Russia: https://tinyurl.com/y7pfwhhu
[6]US spy operation that manipulates social media: https://tinyurl.com/jftugvn
[7]Operation Earnest Voice (wikipedia): https://tinyurl.com/q8egn3z
[8]Twitter (2020): https://tinyurl.com/ybwq4fau

Although the practice of targeted disinformation is not a novelty, it was only after the events surrounding the Facebook-Cambridge Analytica (CA) scandal[9] that the use and effectiveness of such mass 'manipulation' tools began to attract public attention. Given the sensitivity and importance of the application domain in which the above mentioned company operated, several electoral events passed under the magnifying glass suspecting external interventions to influence the results. For instance: the Brexit referendum [80], the US Presidential election in 2016 [17], the elections in France [60], Mexico [21], Kenya [112]. CA's executives stated that *"Cambridge Analytica and its parent company, Strategic Communications Laboratories (SCL), have worked in more than 200 elections across the world, including Nigeria, Kenya, the Czech Republic, India and Argentina"*[10]. These are few well-known examples of how and to what extent bots and disinformation can damage democracy, with the risk of leading to diplomatic affairs between nations[11] [119].
Unfortunately, the effectiveness of conducting disinformation against a targeted audience is not limited to the electoral, political or governmental field. Even sectors such as the economy [14, 59], climate change [113], human rights [140] and, especially at present time, health (e.g., anti-vaccine movements [27, 87] and currently COVID19 [25, 35, 131]), are not spared by campaigns of disinformation.

Governments[12], academics [187] and OSM administrators[13,14] are struggling to control these problems. Although studies for fake news detectors are reaching remarkable objectives, the detection of such malicious content still remains an open problem in the scientific community [38, 152]. Moreover, due to the inaccuracy of the current fake news detectors [151] and the risk of leading to censorship, the most widely used approach for avoiding misinformation still consists in identifying malicious accounts, bots in particular. Despite the efforts spent by OSM administrators in removing suspicious accounts[15,16,17], and by researchers in improving bot detection techniques [4, 118], this plague is far from being eradicated. In fact, according to a 2017 estimate, there were 23 million bots on Twitter (around 8.5% of all Twitter accounts, in [166] they are estimated to range between 9% and 15%), 140 million bots on Facebook (up to 5.5% of all Facebook accounts) and approximately 27 million bots on Instagram (8.2% of all Instagram accounts)[18]. This has motivated a vast body of work on bot recognition in

[9] Facebook–Cambridge Analytica data scandal (wikipedia): https://tinyurl.com/y9rorxln
[10] Cambridge Analytica: The data firm's global influence (BBC):https://tinyurl.com/yxlo8f3u
[11] U.S. Accuses Russian Military Hackers of Attack on Email Servers (source The New York Times): https://tinyurl.com/yae3k674
[12] https://tinyurl.com/yym2xa3v
[13] Facebook: https://tinyurl.com/yac7lsn6
[14] Twitter: https://tinyurl.com/ybx5tn4o
[15] Facebook (2019): https://tinyurl.com/y3yzvpah
[16] Twitter: https://tinyurl.com/y3efs8s5
[17] Twitter (2020): https://tinyurl.com/ybwq4fau
[18] Combatting Targeted Disinformation Campaigns: https://tinyurl.com/ybr4ntw2

social media [29, 68, 155, 160]. In particular, recent approaches to bot detection on Twitter rely on directly observing specific features, such as the ratio of friends over followers of a registered user, the quantity or frequency of their interactions, the expressiveness of their comments, the presence of a name, face photo, address, biography or any additional information on the profile [34, 44, 48, 166]. Although bot detection is undoubtedly effective in fighting misinformation, the risk of being faced with the hydra effect is real. In fact, also due to the cheapness and legitimacy of bots' purchasing and selling[19], when some of them get detected, and then suspended or removed, a new generation of more sophisticated ones can come into play with the ability of avoiding detection (even updating those bots who survived to detection) [47].

The role of human beings in this field does not seem to have received as much attention. Indeed, most of the work in the literature dealing with fighting fakes and mis-/dis- information is focused on studying their effects. the best of our knowledge, just a few of them [116, 169, 170] are aimed at facing the challenging problem of proactively identifying fakes. More specifically, in [169], the authors investigate the features of genuine users starting to interact with a social bot. Instead, in [170] the authors addressed the problem of singling out those features useful to predict whether a user is likely to interact with a bot. Finally, in [116] a comprehensive categorisation scheme for social bot attacks in Twitter has been proposed by modelling the different attack dimensions (i.e., *targets, account types, vulnerabilities, attack methods* and *results*) claiming evidence about the impact of social bots in link creation between targeted human-operated accounts in Twitter.

In this thesis we devote our attention to those human users on OSM which follow a relatively large number of bots (*bot-followees*); by abuse of language, we call those human-operated accounts *credulous* users. In our opinion, it makes sense to think that the risk of being exposed to deceptive content (e.g., fake news) increases proportionally to the number of (potential) malicious entities, such as bots and bot networks, a user is following. From this perspective, credulous users can be an easy prey for mis-/dis-information campaigns, especially those targeted to a selected audience (*targeted disinformation*). For the purposes of this thesis work, we do not consider essential the evaluation of the intent for which (false) information is generated; therefore, the concepts of mis-/dis-information have to be considered exclusively as a consequence of the activities of credulous users on OSM.

With the aim to improve users' awareness of the threats arising from what users read, disseminate and believe to be true on OSM, we will work out an approach capable of singling out credulous users, supported by Machine Learning (ML) techniques. Then,

[19] Compra-seguidores: https://tinyurl.com/y9hs482s

we will thus investigate the behaviour of credulous users by focusing on the actions they perform on OSM. In particular, we will analyse the involvement of credulous users in supporting potential harmful activities, for instance, by bouncing bot-originated content and/or disseminating fake news.

Given the importance of the addressed issues, several stakeholders (e.g., OSM administrators, governments, academicians, etc.) can benefit from our findings. For instance, progresses in this topic can be useful to: (i) identify potential targets of mis-/dis- information campaigns in advance, (ii) protect human users from attacks performed by malicious OSM entities (e.g., bots), (iii) limit mis- /dis- information phenomena on OSM, (iv) increase usefulness, credibility and effectiveness of social media (hence also the content published in such platforms), and (v) safeguard democracy.

RESEARCH QUESTIONS. To achieve our goals, the following research questions have been formulated, and they drive the investigation and analysis performed in the following chapters.

The suspects about the role of bots in polarization phenomena of human users through dis-/mis- information, have led us to direct our attention to those human users (namely credulous) following many automated accounts on OSM. This may be due to the inability of some users to distinguish human from automated accounts; we refer to this inability with the term of users' gullibility. Our first research question is:

> **RQ1** – Among human Twitter users, which type of social relationship (e.g., following or being followed) is the most influential, and why? Does it make sense to assign a *gullibility* score to human users? Which user-related aspects should be taken into account in such a score? Does a clear separation between credulous and not credulous users exist? Or, simply, is one user more credulous than another? (see Chapter 3)

Reasonably, the identification of credulous users implies inspecting the (many potential) contacts the users are following. To avoid this computationally expensive task, we ask our second research question:

> **RQ2** – How effectively Machine Learning (ML) techniques can be in distinguishing credulous and non-credulous users? Is it possible to avoid in depth inspection of human users' social contacts in order to lighten the complexity of identifying credulous users? What is the loss in terms of accuracy when performing their identification? What are the features of Twitter accounts that can facilitate this distinction? Are the features used for bot detection beneficial also for identifying credulous users? (see Chapter 4)

Although on the one hand, the automatic identification of users in social relationships with a considerable number of bots (credulous users) is important, on the other hand, knowing (or at least trying to estimate) the amount of bots a human user follows is definitely valuable. This brings to the third research question.

> **RQ3** – Is it possible to predict the number of bots a human user is following (*bot-followees*)? Are the features, used for credulous classification, useful also for this task? Which measures can be adopted to estimate the quality of such predictions in absence of well-defined benchmarks in the literature? (see Chapter 5)

As further investigation, we started to understand whether credulous users actually behave differently (in terms of actions performed on OSM) w.r.t. not credulous users. Besides, it would be valuable to investigate the level of involvement that credulous users have in bouncing (potentially malicious) content produced by bots. Finding an answer to the following fourth research question should shed light on these aspects.

> **RQ4** – Is it enough to compare the different types of activities (i.e., retweets, quoted tweets, replies and posting new content) between credulous and not credulous users to significantly differentiate them? Can bot-followees influence, in terms of content production, the activities of credulous users more than not credulous ones? How to measure the effectiveness of such an influence? Do credulous users bounce bots' content? And to what extent with respect to not credulous users? (see Chapter 6)

Investigating how and to what extent credulous users bounce content produced by bots undoubtedly provides useful information about their level of involvement in supporting potential malicious bots activities. To understand if credulous users contribute to spread fake content, we analysed those users who posted news, through their favourite social media channels, whose fake nature is undoubted. To this end, we aim to provide an answer to our fifth research question.

> **RQ5** – Do credulous users contribute to fake news spreading? What is their level of involvement compared to that of not credulous users and bots? Is it possible to provide evidence of credulous users contributing to misinformation? Can we take advantage of credulous users detection for fake news detection? (see Chapter 7)

CONTRIBUTION. The main contributions of this thesis can be summarised as follows:

1. we provided an approach to rank human-operated accounts by measuring their gullibility, i.e., by using introducing some heuristic rules (e.g., seniority on OSM), and a method to single out an initial list of credulous users;

2. we refined the credulous users' identification process by training some decision models (by means of ML algorithms) to distinguish credulous from not credulous users among human-operated accounts;

3. we conducted a feature analysis investigation to determine which are the most discriminant ones for credulous users using as a stating point the feature sets defined in literature [44, 49];

4. we generalized the credulous detection approach to all humans by training regression models capable of estimating the percentage of bots that a human user is following;

5. we conducted a behavioural analysis to investigate the activities of users in terms of actions performed on their dashboard, with the goal of understanding the differences between the users identified being credulous from those being not;

6. taking into account the source of the content bounced by credulous and not credulous users, we provided evidence that the former is more prone to diffuse potentially misleading (or anyway unreliable) content. We carried out statistical tests to check the significance of these results, and to reinforce this evidence;

7. we investigated the harmfulness of credulous users, always compared to not credulous ones, in terms of fake news dissemination. We performed the analysis by focusing on: (i) the number of fake news, (ii) the number of published tweets (corresponding to fake news), and (iii) the amount of users involved in their spreading;

8. we provided some datasets of credulous and not credulous users[20] that are publicly available to the scientific community.

Thesis Organisation

The remainder of this thesis is organised as follows. Chapter 2 provides the background knowledge; precisely, Social Media Mining, Misinformation (in all of its forms), and ML are explained to furnish the reader with some concepts that can be useful to better understand the investigations, the experiments and the analysis performed in this thesis. Chapter 3 discusses how to identify credulous users and which aspects may be useful for their categorisation with respect to other human users. Chapter 4 exposes a study

[20] Credulous users datasets: https://github.com/AlessandroBalestrucci/Credulous

on a larger dataset of humans (w.r.t. Chapter 3). We try to automate the recognition process of credulous users by means of learning algorithms building decision models. Further, we try to figure out which features distinguish credulous users best. Chapter 5 generalises the problem of identifying credulous users by extending the study to all human-operated accounts. The goal is build predictive models that quantify how many bots (in percentage terms) are infiltrated among the social friends of human-operated accounts. Chapter 6 deals with behavioural analysis conducted on credulous and not credulous users, to find differences between these two typologies of human-operated accounts. In particular, two kinds of analysis were conducted at a different level of detail. The first analysis takes into account only the actions performed by users (in terms of posting types). The second analysis focuses on the authors of content bounced by humans, to establish the actual involvement of credulous users in spreading content originated from bots. Chapter 7 exposes our most relevant findings. We provide evidence about the dangerousness of credulous users as active entities on OSM when spreading and amplifying disinformation. Chapter 8 conclude this work by resuming the main findings from our investigation, some possible practical usage and future research directions.

Chapter 2

Background

This chapter aims to provide the reader with the basic concepts and knowledge useful for a better understanding about the topic, main concepts and experiments which the research of this doctoral thesis.

Section 2.1 introduces online social media (OSM). In Section 2.1.1, an overview of the research field called *Social Media Mining* (SMM) is provided. In Section 2.1.2 it will be explained what is *Social Media Analytics* (SMA) and the differences with SMM. The intent is to better frame the research direction decided for the next chapters of this thesis. A brief overview on the typology of the different OSM will be provided, by mainly focusing on the description of our operational context, i.e., *Twitter* in Section 2.1.3. More in detail, it will be presented which are the basic Twitter features, the type of actions and relationship between users in such media.

Section 2.2 discusses the principal domains (like business, public security and political communication) where the analysis of Social Media data represent a key factor, and consequently, the research is very active and important. In Section 2.2.1, special attention will be given in exposing what is defined as the main problem on social media, i.e., the *misinformation*. Specifically, what misinformation is, which forms can assume, who is responsible of this phenomenon and how it spreads over the social media.

Section 2.3 provides a brief introduction on Machine Learning (ML) and its sub-areas, detailing more those concepts useful for understanding the performed experiments. A brief description will be given on the machine learning techniques most used in SMA, and more specifically in misinformation fighting. This chapter ends by describing the ML tasks needed for this research, the followed experimental methodology and the tools that have been used to achieve the planned objectives.

2.1 Analysis on social media

The rise of social media in the digital world has increased and certainly revolutionised the ability to socialise and interact among people [120, 161]. Since their first appearance, dating back to the last decade of the previous millennium, social media have been positively appreciated by people, especially by youngsters[1]. Year by year, users of all ages and from everywhere started to get increasingly closer to these social interaction platforms by routinely interacting with their friends. Initially, Social Media had the objective of fostering and facilitating communication and sociality among users, interconnecting them with each other regardless of geographical distances. In fact, to define such web platforms for people interconnection, up to not so much time ago, the term Social Network sounded more common than Social Media. But nowadays, especially among non-expert users, the difference between these two terminologies is negligible and used as synonyms[2]. It is worth to say that, from a technical point of view, social media and social network do not identify the same concept. Indeed, the term social media refers to that specific set of technologies of Web 2.0 that allow users to generate content and establish social relationships between them; forming the social network. Therefore, we can say that the social network can be built thanks to the social media[3]. Furthermore, the representation of the social network depends by the types of relationship may occur between the users; e.g., in most of cases, a real social network (consisting of strong ties among users) or of a so-called conversation graph (especially in the case of social platforms where relationships are built around discussion topics of interest). The capability to support users' interconnection has undoubtedly been the strongest point of social networks, but being tied to a single strategy would not allow such web platforms to survive so long in the digital world. Let us think to the evolution from *FaceMash* to *TheFacebook*, e.g., to the delayed introduction of the *Wall* and its further evolution in *Timeline* [4] [134]. The key factor, that boosted the usage and the importance of social networks, has been to stimulate their subscribers to produce content (user-generated content), e.g., the introduction of hashtags in Twitter (one year after its launch)[5]. In such a way, social network began to be considered a means of communication like tv, radio or newspaper, and in many cases more effective of the other media [122, 123]. This feature, maybe not too much considered previously, has become the strongest point of these platforms. This gave them greater visibility, but also enlarged the range of users. This was also beneficial for companies that can exploit social media in their market strategies, thus to widen the audience and reduce advertising costs. The increasing

[1] Facebook: https://tinyurl.com/ycm8xomn
[2] Powerthesaurus: https://tinyurl.com/y7b3aq7x
[3] Social Media and Social Network differences: https://tinyurl.com/y6eko2sq
[4] Facebook's features: https://tinyurl.com/yxuaw2qt
[5] Twitter hashtag: https://tinyurl.com/y72vjjtl

number of active users, which social media can boast of, is in the order of billions. It is obvious that such a multitude of users produce an uncountable amount of data, from which it is possible to derive useful information [176]. The burden of data crawling (from the social platforms) falls to a research field called *Social Media Mining* (SMM); whereas, the transformation of data into information concerns to another research field called *Social Media Analytics* (SMA), strongly connected with SMM [157, 185].

2.1.1 Social Media Mining (SMM)

Social media mining is an emergent research field, whose main task is to obtain User-Generated Content (UGC) resulting from social interaction (e.g., posts, retweets, replies, etc.) on Online Social Media [185]. What differentiates Social Media Mining from other forms of mining, such as data-, text-, web- mining, is the peculiarities of the data circulating on Social Media. First of all, with respect to what is generally supposed in data mining, it is not possible to assume that in OSM the data respect the *i.i.d.* property (independent and identically distributed [36]) because: (i) in most of the cases, data show a power-law distribution [185] and (ii) they can be dependent/linked each other, e.g., replies to other posts or content in pages/channels/lists. Another aspect of Social Media data is what can be called as content's *sociality*. This concept has nothing to do with concepts such as *Social Big Data* [73] or *Social Media Big Data* [107], whose issues mainly concern problems of Big Data and hence, at least in part, outside the scope of this work. Usually, it is common to mainly divide the data in two categories [7]: *structured* data, i.e., highly-organised and formatted (e.g., data in relational databases), and *unstructured* data, i.e., without pre-defined format or organisation (e.g., text). In Social Media there exist both these categories; for instance: structured data may be relational data, explaining the relationship between users (e.g., mutual friendships, followers, followees, etc.), unstructured data may be users' text (e.g., posts, tweets, comments, replies, etc.). Furthermore, users can add another kind of data that falls under the definition of *multimedia* data (from the web mining field); such as: web site url(s), pictures, videos and vocal messages. But what differentiates the data in Social Media from the others, it is the possibility to generate content, through "keywords/key symbols", e.g., hashtags (#), mentions/tags (@), and so on, with links to other OSM entities such as other users, pages, channels, etc. We refer to this peculiarity as *sociality* of data. Due the plethora of data types and *sociality* existing in OSM, it is mandatory to consider Social Media Mining an *interdisciplinary* research field, whose methods and theories encompass among multiple fields such as social science, ethnography, statistics, machine learning, social network analysis, mathematics, and statistics [185]. Moreover,

from these peculiarities, Zafarani et al. derived some challenges that Social Media Mining has to deal with, grouped in four tasks/definitions: *Big Data Paradox*, *Obtaining Sufficient Samples*, *Noise Removal Fallacy* and *Evaluation Dilemma* [185].

Big Data Paradox. Although at first glance the amount of data on Social Media may seem abnormal (and indeed they are), from an operational point of view this is not really true. In fact, the mining phase is preventively designed to answer certain research questions concerning a very restricted context and population among OSM's entities. However, this sort of filtering can also result in a very large amount of data to process/elaborate, but much smaller with respect to consider the entire Social Media system. SMM faces this issue by exploiting the structure of social media and its multidimensional, multisource, and multisite data, information is aggregated to increase the efficiency and effectiveness of the mining phase.

Obtaining Sufficient Samples. To automatically collect data from OSM, Application Programming Interfaces are made publicly available by the provider of social platforms. But, despite this method makes the download easy and convenient, some constraints and limits are posed by OSM administrators to protect their data, the privacy of their users and, of course, for cybersecurity issues (like DDOS attacks). About the mining, both researches and developers face with a common problem: only a limited amount of data can be downloaded (e.g., per day or number of requests). Due to such restrictions and the missing knowledge about OSM users's distribution, SMM has also to verify that the amount of downloaded samples are sufficiently representative of the social population data, or at least of that portion of users under investigation. This is a very important issue to guarantee the robustness of the findings resulting from related data analysis.

Noise Removal Fallacy. In several machine learning applications it is normal to find fallacy data, known as *noise* (e.g., outliers). Noise removal is an important step in data preprocessing phase to increase the quality of datasets and, consequently, the reliability of subsequent findings. In SMM this process is complex due to the nature of social media data, they can include a not negligible quantity of noise. The removal of noise, especially without well defined criteria, can led to flimsy experimental scenarios, because the removal process can also delete data that, despite assumed as noisy, can contain valuable information. Another issue is given by the definition of *noising data* because it is relative and strongly dependent by the domain of application, the problem and the operational context.

Evaluation Dilemma. As well known, the starting point of the Knowledge Discovery in Databases (KDD) process is some kind of *ground-truth*, mainly constituted by a dataset from which it is possible to derive new information under the

form of patterns in data. Then, the dataset is divided in two parts: *train* and *test* sets. Roughly, learning algorithms are executed by considering training data only to produce decision/clustering models. After, the models' performance are evaluated by means of the *test* set, to measure the reliability of such models with unseen data. Sometimes, the ground-truth is divided in three parts: *train*, *evaluation* and *test* sets. This happens when the learning process needs to be iterated multiple times [117]; for instance, in hyper-parameter tuning (called *iteration*) and deep learning (epochs). The learning is always performed on the train data and, at the end of each iteration, the models' performance are evaluated with the *evaluation* data that acts as an "online" test set. Then, learning algorithm's parameters are tuned and the learning phase is repeated. When the trained model's performance achieves good scores, such model is finally tested with test data. In SMM, the availability of a ground-truth is not a foregone. This makes *Evaluation Dilemma*, perhaps, the most challenging task. The public availability of ground-truth in SMM strongly depends on: (i) the domain, (ii) the data availability, and mostly (iii) the advances in the state of the art (w.r.t. the context and problem).

2.1.2 Social Media Analytics (SMA)

In the literature, there are several definitions about Social Media Analytics (SMA). In [186], SMA is defined as *"an emerging interdisciplinary research field that aims on combining, extending, and adapting methods for analysis of social media data"*. In [156], SMA is considered as *"the art and science of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making"*. In [157], SMA is described as *"as an emerging interdisciplinary research field [...] for gathering, modeling, analysing, and mining large-scale social media data in order to make business, economic, social, and technical claims from both research and practical perspectives"*. To generalise the approaches for different domains and purposes, in [157] SMA is modelled as a *process* composed by four steps (see Figure 2.1): *Discovery*, *Tracking*, *Preparation* and *Analysis*.

In the *discovery* phase, it is set the domain and the problem to face by collecting all the related information and the knowledge from the state of the art. This way, it is possible to restrict the investigation by focusing on the most relevant data only.

In the *tracking* phase, it is decided: the social platform (i.e., the source of data), the data crawling methods (e.g., APIs), the approach to query the OSM and how to store the downloaded data (w.r.t. the type of data).

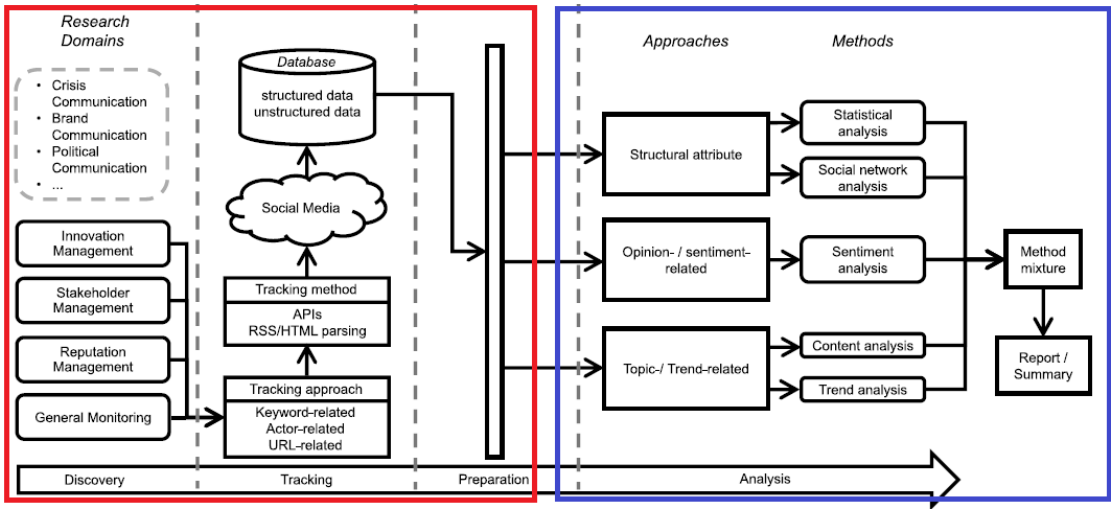


FIGURE 2.1: Social Media Analytics processes as described in [157]. In the left box (red), there are the steps we considered belonging to Social Media Mining field. In the right box (blue), the issues related to Social Media Analytics tasks.

In the *preparation* phase, problems related to data consistency, quality and reliability have to be addressed. In particular such problems concern how to deal with noisy or incomplete data that can affect the obtained findings.

In Figure 2.1, these three steps are grouped together because, from what we learnt about Social Media Mining in Section 2.1.1, they mainly overlap with the aims and challenges of Social Media Mining.

The last step of SMA process is called *Analysis* (see Figure 2.1), and we can consider it as the core step of the whole SMA process. In this phase, with respect to the objectives and the domain, it is decided the approach to achieve the goals (e.g., community detection, sentiment/opinion discovery, trend investigation) and the proper method (or more than one, *method mixture*) of analysis. This way, it is possible to produce new and non trivial knowledge that may be quickly used (e.g., in decision making processes).

From the several definition of SMA in literature, two are the concepts on which there is an agreement: to consider SMA as an (i) interdisciplinary field and (ii) the essentiality of this process to derive new and non trivial knowledge from the analysis of Social Media Data. On the other hand, what differentiates these definitions is how many and what tasks the SMA has to accomplish before performing the data analysis.
Summarizing, SMM and SMA can be considered as two distinct but symbiotic research areas. SMM concerns data retrieval and management, SMA concerns approaches that allow to extract knowledge from data (obtained via SMM).

2.1.3 Operational context: Twitter

In literature [3, 86, 185] social media are divided into categories according to certain criteria that take into account (but not limited to): the type of content managed (user-generated content), the type of (social) relationships (between users), the interaction's capability.
Some of the most well-known social media categories are detailed below:

Blogs: websites where the published content appears in a chronological fashion and the signed users/visitors can read and comment (e.g., The Huffington Post);

Microblogs: similar to blogs but with constraints on the length of posting. Users can also keep up with news from other users (e.g., Twitter);

Collaborative projects: platforms that foster and stimulate cooperation between users with different skills/knowledge, but who share has the objective of developing a public utility project and making it available to the public (e.g., Wikipedia);

Social networks: the interactions between users derive from a real mutual knowledge existing from the real world or the sharing of interests such as participation/organisation in events. Such platforms can incorporate features of photo/video sharing and instant messaging (e.g., Facebook and LinkedIn);

Products/services review: platforms where users share/publish their reviews about products and services and the reliability of the shops/sellers (e.g., Amazon, Ebay, Tripadvisor);

Photo/Video Sharing: websites that offer services such as uploading, hosting, managing and sharing of photos and videos (e.g., Instagram, YouTube, TikTok);

Social gaming: online gaming platforms requesting a (social) interaction between the players/users (e.g. World of Warcraft);

Virtual worlds: online digital environment where the signed users can impersonate an avatar, explore places and interact/socialise with others (e.g., Second Life).

In this research we focused on *Twitter*, a popular social networking service on which the interaction and the posting activities, among users, is via *tweet*(s). A *tweet* is a message that can contain text, links, visuals or a mixture of them. There is a limit for the tweet's length of 280 characters (before it was 140). This OSM allows to unregistered users to read tweets only, registered users instead have access to the complete publishing features as: post/tweeting, like, and retweet/quote other tweets. People make connections by following other people's twitter feeds. Once a user clicks the *follow* button in the main

page of another user, anything that person or organisation tweets will appear on its timeline.
The reasons behind our choice to consider Twitter as a benchmark are basically three: (i) the brevity of the content posted, which therefore does not tire/yearn users in reading (in fact, this feature is considered to be the most successful factor of microblogging platforms), (ii) the ways in which connections are established between users, in fact to access content published by a user is not necessary a reciprocity of the relationship (e.g., in Facebook); and, (iii) an ease in downloading data from the platform and dataset availability in the literature.

The kinds of tweet and social relationship

Tweet. A message posted by a registered user containing text, photos, a GIF, and/or video. It is necessary to insert/link the content and push the **tweet** button (Figure 2.2).

FIGURE 2.2: Tweet.

Retweet. A tweet published by other users and automatically shared on their timeline. In Figure 2.3, it is possible to see two retweeting modes. The first immediately shares the original tweet without modifications. The second allows to make a quoted (tweet).

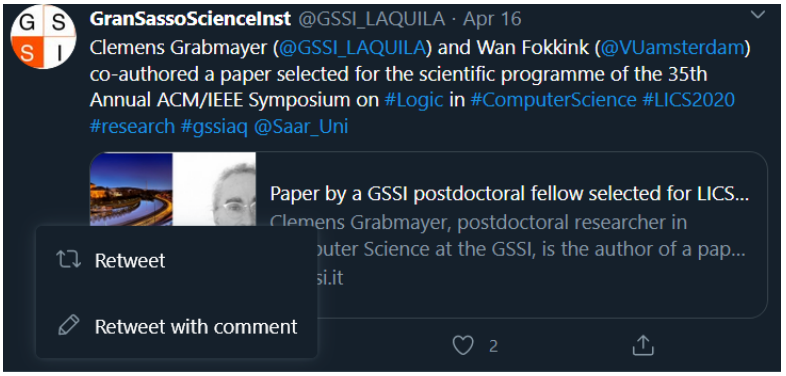


FIGURE 2.3: Retweet modes.

Quote. It is a retweet with own added comment to the tweet another person's tweet. In Figure 2.3 it is possible to see the sentence *Retweet with comment* which refers to this kind of tweet. In Figure 2.4 it is reported the button in such case reports the text *retweet* as a confirmation of a particular type of retweet.

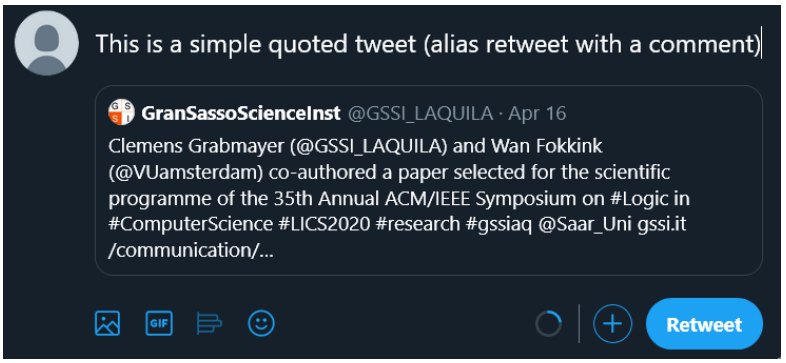


FIGURE 2.4: Quote (tweet).

Reply. It is a tweet to respond to another person's tweet. In Figure 2.5 it is possible to see that on the button it is written *reply*.

FIGURE 2.5: Reply.

Mention. It is a tweet containing another account's Twitter username, preceded by the "@" symbol, see Figure 2.6.

FIGURE 2.6: Mention (tweet).

Social Relationships: *Followers* and *Followees*. As mentioned before, social media also differ from each other by the way they represent/implement their own social

FIGURE 2.7: Followers and followees counters

network. A social network is built through the connections that users establish between each other and these connections are developed by following certain types of relationships that exist between users. In general, as explained in [72], these relationships can be generated on strong ties (e.g., friendships that from real life are transferred to the digital platform) or weak ties (e.g., users who do not know each other but share interests).

In Twitter there are two types of social relationships among users: *followers* and *followees*. The *followers* are those users that are interested about the update of a certain account, and all the tweets she/he produces will be shown on the followers' main page (namely *home timeline*). This way, a Twitter account shows its interest to follow a user's activity in the social. On the other hand, the *followees* relation expresses the interest of a user to receive, in its own *home timeline*, all the tweets produced by the accounts she/he is *following* (in Twitter web site, the word *following* is used to define the followees). In Figure 2.7 it is possible to see the amount of *followers* (right box) and *followees* (left box, namely *following*). From this we can define that the followees relationship is the one that most expresses the concept of interest in something; on the other hand, there would be no followers if someone did not decide to follow someone. Moreover, the Twitter recommendation system only suggests (logically) whom to follow (i.e., *followees*); therefore, although strong links may take place on this social media, we believe that the social network in Twitter mainly bases on weak ties [16, 135]. That is not a disadvantage, as one of the main findings in [72] is that "in marketing, information science, or politics, weak ties enable reaching populations and audiences that are not accessible via strong ties". Due to its semantic meaning and properties, the *followees* relationship is a key concept in this research work.

2.2 Research in social media

Taking into account the amount of audience and the high impact on the real world, in [157], among several topics and domains treated in Social Media's discussions, three have been identified as the most important ones: (i) business, (ii) public security and (iii) political communication. Governments, academicians and OSM administration are the principal stakeholders on the related research that becomes increasingly topical.

The research related to the *business* domain mainly concerns about private companies for purposes related to the improvement of their business. In particular, for this kind of stakeholders, the analysis of Social Media, and related investigations, are useful for several reasons: to check and improve the company's external reputation (e.g., by protecting against bad advertisements, improving the company's public image) [32], to detect the rising of new trends, to improve their marketing campaigns and place new products [13], and to improve the communication with their customers [53], especially in the case when they need assistance.

For what concerns research in *public security* domain, the stakeholders are individuated in the governments and public organizations/agencies. A smart usage of Social Media during emergency situations, as happened in several recent events, can represent a very important factor to better guarantee the safety of people, especially during a situation of crisis (e.g., natural disaster). The data obtained from Social Media (e.g., geolocation, images, videos and so on), if properly analyzed, can provide prompt and previously unknown information when an emergency situation arises. Furthermore, w.r.t. traditional mass media, governments and emergency management agencies can quickly diffuse life-saving information and directives to a wider audience, monitor the current status and the sentiment/opinion of the involved people; e.g., natural disasters [92] and covid-19 pandemic [177].

In *political communication* the main stakeholders are the political parties, governments and communication agencies (both private and public). Especially during election campaigns, people discuss on Social Media their concerns, proposing actions and solutions that should be taken into accounts from politicians and/or governments. By analysing such data, political parties can easily reach a wider audience and perform an ad-hoc political campaign in order to gain consensus among the active electorate, hence to design a smarter political campaign. By using their official social channels, both governments and politicians can define better strategies to gain more visibility and, consequently, more followers on their political opinions especially increasing the interaction with them. Text/opinion mining joint with social network analysis are the principal methods with which

to derive important findings and to improve the efficiency and effectiveness of a political campaign , e.g., US election in 2016 [179]).

Another important line of research, recently emerged because of its negative effects, is what has been defined in [129] as "the dark side of social media". With this term, the phenomenon of misinformation is referred, it damages not only the reliability of OSM as information sources, but it also affects the users attempting to change their opinion or generate misconception. This increasingly widespread phenomenon is caused by the production and dissemination of unreliable and/or misleading content, known as *fake news*. The dangers and effects misinformation produces in the real world are widely recognized by academics, governments and OSM administrators, who are constantly struggling to face this issue. Brexit and the US Presidential election in 2016 are just some of the most studied scenarios recently under investigation by scientists. For these reasons, methods and approaches for effective and quick detection of fake news are a hot and challenging topic in the scientific community research.

2.2.1 Mis-/Dis-information and spreaders in OSM

The result of mis-/dis-information containment strategies strictly depends on the effectiveness of intercepting and eradicating fake news on OSM. In [94], *fake news* are defined as all those information that are untrue, fabricated or in any case unverified/supported by a fact checking process to ensure their accuracy and credibility, e.g., clickbaits, hoaxes, rumors, urban legends and so on. In [173, 174], a categorization of fake news has been proposed according to the increasing degree of intentionality to deceive/polarize the users, as reported hereafter.

1. *Satire or parody.* Category of news created just for entertainment purposes to stimulate humour in the readers. With this kind of news, there are no harmful goals, but for those who do not understand the satirical purposes, they have the potential to deceive, especially if the news look like a reliable source. This can happen when the reader has low confidence with the treated topic. Examples are web sites like *Lercio* or *Butac* (in Italy) or *The Onion* (in USA);

2. *False connection.* News where headlines, visuals or captions do not support the content. The main goal, about their production and dissemination, is to attract user's clicks for profits, obtained by the advertisements shown in the redirected web sites. When people start reading the full article, they realize of being deceived. A well-known example of this type of content is *Clickbait* headlines. For instance, on Facebook is know the usage of visuals or captions to this end, i.e., users scroll them without clicking but headlines can be deceptive;

3. *Misleading content.* When news production is engineered by making misleading use of information to frame problems or attack individuals, aiming to attract (often) public support. Cutting photos, choosing quotations or selecting statistics wisely, are just some of the methods used by applying what is referred as *Framing Theory* to induce a bias in readers' perception [55]; This kind of content is popular especially when politicians want to decrease the credibility of their opponents; images are the favourite mean because of their immediateness;

4. *False context.* The content of the news is genuine but not the context to which it refers. Examples are the use of images of disasters (from different periods or places) referring to current situations, or more generally, to a different context;

5. *Imposter content.* Content created from a fictitious source with characteristics (e.g., logo, site name and themes) very similar to those of a reliable one and attempting to "impersonate" it;

6. *Manipulated content.* Original contents (e.g., stories or imagery) are manipulated for deception purposes. *Half-truth* can be attributed to this category;

7. *Fabricated content.* News ad-hoc fabricated and completely false, created and disseminated to deceive/polarize people and do harm.

By considering the intentionality (of fake news) to deceive (or not), it is possible to distinguish disinformation from misinformation [174].

Misinformation includes all those fake news information disseminated by people without harmful intention. According to the previous fake news typology, *satire/parody, false connection* and *misleading* content fall in this category. As opposite, *disinformation* refers to all false information produced and spread with harmful intent, like to induce bias in public opinion, hate speeches, to generate misconception and so on. *False context, Imposter-, Manipulated-* and *Fabricated-* content fall into this category and is widely used especially during political campaigns. Furthermore, in [174], it is also exposed the concept of *malinformation* as the leakage, and subsequent dissemination of "genuine" information to cause image damage and increase discontent. In this category fall some data leaks, harassment and hate speech. Recent events are the Hillary Clinton's emails leakage, during US presidential election in 2016, and Macron's emails leakage in 2017.

To increase the effectiveness of misinformation, fake news need to be disseminated as much as possible, and this is pursued by some OSM users. Mainly, there are two categories of users: human-operated and automated accounts, also known as *bots*. Human-operated accounts, as the name suggests, are all those OSM accounts managed by real persons. As opposite, *bots* are those accounts managed instead by ad-hoc software able

to perform actions in OSM. Accordingly to what they do and the extent of their interaction capability, several types of bots exist in OSM [71]:

Web robots are those automated accounts known as *crawlers* and *scrapers*. They do not interact with users but are mostly used to automatize the download of Social Media data to obtain datasets for SMM or SMA puroposes;

Chatbots [50] are entities in OSM (but not only there) able to interact through simple natural language (text or speech). They can be defined as dialogue systems with a lexical interface. Vocal assistants like Apple's Siri and Amazon's Alexa are examples of the highest evolution of this category of bots, they are widely used also by companies and public offices to improve their consulting services (e.g., automatic response);

Spambots [18] show clear malicious purposes like viral marketing, private information theft, harassment of legitimate users and so on. Their principal activity is to catch the attention of users, by increasing their visibility, to bomb them with unreliable information transmitted by private instant messaging, posting often using links, visuals or videos, email and so on;

Sockpuppets [15, 26] are those accounts that in disguise perform malicious activities such as expressing provocative opinions, targeting individuals, manipulating public opinion and spreading misinformation by using a posting protocol;

Social bots [61], also called *sybils*, are OSM accounts controlled by ad-hoc software applications, able to mimic the behaviour of human users (in term of content production activities). Sometime they are used for benign purposes, but most of the cases they are created (and used) to harm other users (especially, the humans) in order to manipulate, deceive and/or polarize them. Due to their interaction capability to relate with genuine users, they are widely applied for malicious goals like stealing personal data, driving political conversations, manipulating marketing campaigns, and especially spreading mis- and dis-information.

About human-operated accounts, a categorization is indeed not a trivial task. But, looking at the interaction that human accounts have with malicious entities (e.g., bots or fake followers [44]), and/or deceptive content on OSM, we can single out the following categories of human/genuine users:

susceptible [169] – interact with social bots giving, in some way, visibility to bots' activities and cooperating with them;

gullible [149] – accounts replying to known fake news;

- ***credulous*** [11] – follow a considerable amount of bots in Twitter (called *bot-followees*).

A further category of OSM users is represented by *hybrid-accounts* or *cyborgs* [33]. It includes those accounts showing characteristics of both automation, typical of bots, and human behavior. The main issue, in dealing with this kind of users, concerns how to quantify their level of automation so that such users are recognized of being cyborg. Because of this ambiguity, these users are often referred as "bot-assisted humans" or "human-assisted bots".

2.3 Machine learning: a brief overview

The term *machine learning* (ML) have been introduced for the first time by Samuel in 1959 [142]. A formal definition of machine learning, under an operational point of view, has been provided by Mitchell in [115]: "A computer program P is said to learn from *experience* E with respect to some class of *tasks* T and *performance measure* P, if its performance a tasks in T, as measured by P, improves with experience E."

The ML algorithms, starting from a set of data (or a sample), build mathematical models able to intercept patterns in the data. Such models are mostly used to: make predictions, provide support in decisions, identify groups, implement strategies and much more.

Depending on the type of task and the problems being addressed, three basic categories of ML can be identified [141], i.e.:

supervised learning is an automatic learning technique that, starting from a set of observations/instances, whose category/class is known *a priori* (labelled instances), allows to build predictive models. In this category we can mainly identify two types of tasks: *regression*, i.e., the prediction output is a value (integer or real), and *classification*, i.e., the prediction refers to a category/class. With respect to the number of categories we can have *binary* classification tasks, when the instances can be assigned in two classes (e.g., spam *vs.* not spam mail), or *multi-class* tasks, when the number of classes is greater than two. More specifically, in the multi-class task we can identify both single-label and multi-label multi-class approaches [164]. The former approach concerns the instances categorisation into only one of more than two classes; instead, in the latter multi-class problem, there is no constraint on the number of classes that can be assigned to each instance;

unsupervised learning involves all the tasks that deal with *unlabelled* instances aiming to find structures and groups in the set of observations. Its main application is related to *clustering* problems. When the observations in the dataset are partially labelled, we fall into a category called *semi-supervised* learning;

reinforcement learning [162], unlike the two previous categories, concerns with problems of sequential taking decisions, especially when agents have to achieve a specific goal by interacting into an unknown environment or to face opponents. The most common domains of applications, but are not limited to, for this type of learning, are automatic driving systems and the production of strategies (e.g., gaming). The building process, for this kind of models, is pursued through a rewards-based mechanism. Every time the agent does an action useful to reach the goal a reward is given, otherwise, a penalty is assigned.

To extractnew and non-trivial knowledge from data, the well-known process called *Knowledge Discovery in Database* is used [58].

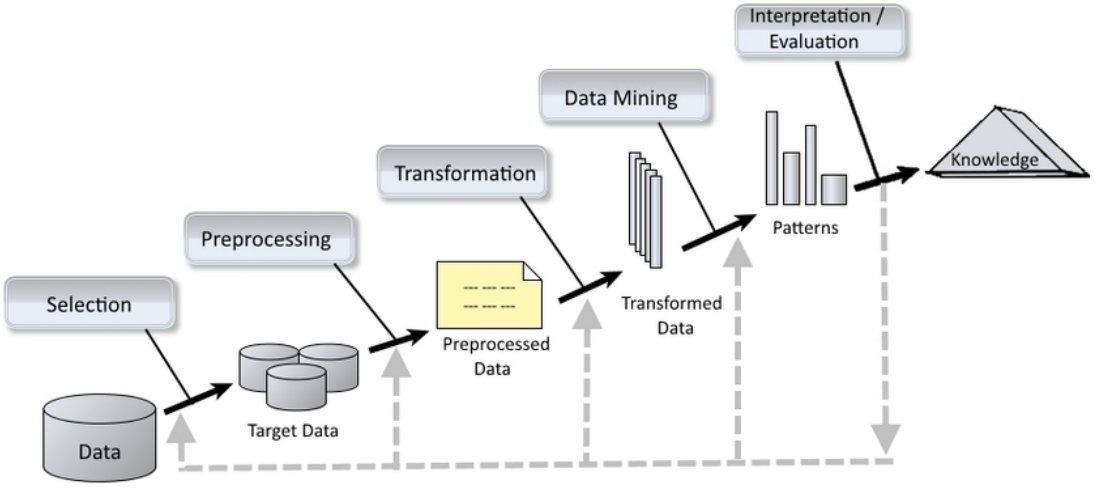


FIGURE 2.8: Knowledge Discovery in Database (KDD) process [58]

As shown in Figure 2.8, the KDD process is mainly constituted by five sequential steps:

Selection – Starting from the whole *data* at our disposal, the first step aims to *select* only those relevant to the type of analysis to conduct, by taking into account the variables to observe and the knowledge we want to derive on. The output is a precise (and reduced) set of data, called *target data*;

Preprocessing – This step deals with all the operations needed to improve the quality of the *target data*; this is possible by removing outliers, handling data-noise (hence, to define the concept of noisy data) and deciding strategies to deal with missing data;

Transformation – This phase involves all the activities referred to *feature engineering* [28]. Accordingly to the specified goal's task, from the raw data are derived the features, through which the entries will be represented (dimensional space of representation). Dimensionality reduction methods (like attribute/feature selection algorithms) are also used to reduce the number of variables to consider for the model training phase. After this step it makes more sense to refer to the entries of our (target) data as *instances*;

Data Mining – Depending on the task at hand (e.g., classification, regression, clustering, etc.), this step concerns about the method(s) selection, hence the learning algorithm(s), to be employed to single out patterns among the instances. At this stage, it takes place the learning process by using a portion of the whole set of instances, called *train* set. The rest of the instances, called *test* set will be used in the next phase to evaluate the trained models. Over the years, various types of ML models have been discovered; the mainly used are: Artificial Neural Networks [127] (specially used in Deep Learning [70]), Decision Trees [138], Support Vector Machines [39], Bayesian Networks [65] and many others;

Interpretation/Evaluation – In this phase, the trained models are evaluated by means of the *test* set, to check how a trained model behaves when previously unknown instances are supplied to it. The model's evaluation strictly depends on the learning task (e.g., classification, regression, clustering); according to the task, different metrics are calculated to quantify the effectiveness of the predictive model. For instance, for tasks' classification the most used evaluation measures are: accuracy (the quantity of instances correctly classified), the precision, the recall, the F1 measure [110], the Matthews Correlation Coefficient (MCC) [111], the Area Under the Receiver Operating Characteristic curve [56] (AUC). This latter measure is very useful in case of unbalanced datasets.

2.3.1 Machine learning and misinformation diffusion in social media

The application of ML techniques to perform analysis on Social Media is wide and mainly intended to derive non-trivial and previously unknown knowledge [1]. One of the most important and more recent challenges is on the spread/limitation of misinformation and fake news. The amount of published work devoted to this challenge via the use of ML techniques is high. To mention few works, although established in the scientific community, would be reductive and would not convey the idea of efforts and achievements in this field. For this reason, we prefer to mention some of the most recent and comprehensive comparative study of current research on the use of ML techniques

for dis-/mis- information detection.

In [41], the authors analysed the problem of mis-/dis-information in the domain of review spam detection. Firstly, they provided an overview of feature engineering performed in several studies from both review-centric (i.e., features obtained by the text) and reviewer-centric (i.e., features derived by the review's author) perspectives. This overview lists all the features that have been mostly used by different works. Furthermore, a discussion on the various ML techniques, proposed for the detection of online review spam, is provided by comparing the corresponding results. Besides, the effectiveness of supervised, unsupervised and semi-supervised methods is discussed. Also in [128], supervised, unsupervised and semi-supervised learning techniques are exhibited and compared in the domain of fake review detection. But, with respect to [41], approaches that, in addition to the features derived from the text or the authors, take into account behavioural and relational features. In [168], the authors provided an overview of approaches proposed in the state-of-the-art to (automatically) assess the credibility in social media. The concept of credibility is described as the quality perceived by individuals to be able to independently discern the genuineness/falsity of information (both sources of information and information itself). Most of the considered approaches in the survey are based on data-driven models, therefore via ML techniques; but, also model-driven and graph-based approaches (especially those works focusing on credibility propagation) are increasingly gaining consideration by scientists. The tasks that mostly concern credibility issues, and on which the authors have carried out this survey, are: (i) detection of opinion spam in review sites (similarly to [41, 128]); (ii) fake news/spam detection in microblogging (e.g., Twitter); and, (iii) assessment of online (both on websites and social media) health information. In their final considerations, the authors point out the lack of predefined benchmarks, gold standard datasets and how the mining of large quantity of data is complicated. Another well-know survey concerning methods to identify fake news is provided in [187]. An interesting novelty of this survey w.r.t. similar works, is the inclusion of multi-disciplinary research works. Furthermore, with respect to other surveys where the fake news detection methods are categorised by looking to the adopted ML techniques or by the usage of social context information, the authors of [187] opt to catalogue the works on fake news detection, considered in their literature review, from four perspectives: *knowledge-based* (i.e., methods based on fact-checking), *style-based* (i.e., which include methods that focus on the analysis of the news content to discover some patterns or writing styles in fakes), *propagation-based* (i.e., methods based on the study of graphs) and *source-based* which mainly deal with work for assessing credibility based on news authors, publishers and social media users. The works included in this latter perspective most match with topic of the previous mentioned work in [168].

Moreover, in [182] three tasks have been identified to better understand and fight the misinformation phenomenon: *diffusion*, *detection* and *intervention*.

In *misinformation diffusion* the focus is on *who* is spreading fake news (called *spreaders*) and in which ways misinformation can spread on OSM. For instance, in this context, ML techniques are used for bot detection. Roughly, a *bot detector* is a decision model that discerns whether an account is automated (hence a bot) or human-operated. Usually, these kinds of models are binary classifiers where the response is calculated based on a feature space through which OSM accounts are represented. For the construction of this type of models it is essential to have a dataset of OSM accounts where the knowledge of who are the bots and humans is known a priori. The quality, quantity, and diversity of data are key factors for an effective bot detector. Another way, to understand misinformation diffusion, is to focus on human spreaders. Here the analysis is a bit more complex due to the heterogeneity of humans' behaviour. With respect to bot detection, the role of human users in misinformation spreading did not receive much attention [11]. Some human users can become the target of bots' malicious activities because, w.r.t. other humans, can be reputed more prone to believe and/or bounce unreliable content (fake news).

Misinformation detection involves all those approaches and techniques able to identify unreliable content such as *fake news*. This category of methods faces the problem of fake news interception in a more *direct* way. The analysis of the content, context and information sources represent the start point to evaluate the truth/reliability of what is circulating on OSM. The most important tasks fall in this category mainly concerns fake news detection. Even in this case, the usage of ML techniques (e.g., text mining, sentiment analysis, semantic analysis, etc.) represent the promising way to pursue this goal [152]. Similarly to a bot detector, a fake news detector is roughly a binary classifier able to judge if news/posts/tweets, given in input, are true or not. Nowadays, fake news detection still remains a very challenging task. The basic idea under fake news detection is to provide, in an automatic fashion, effective fact-checker for an "online" reliability verification of the circulating content.

Misinformation intervention concerns those approaches able to mitigate the spreading of unreliable content on OSM and/or to improve users' awareness of the content they read on the social platforms. Preventive measures like misinformation *-immunization, -isolation* and *-antidotes* are employed to fight against fake news. An example of *misinformation immunization* is to send warnings to specific target people (e.g., those most in danger of misinformation attack by bots). Measures of *misinformation isolation* are the removal or suspension of malicious users on OSM (like bots), and this can slow down the spreading of fake news to general users, thus preventing them from being "infected" [130].

Misinformation antidotes involves easy but, sometimes effective, solutions like official communication by the authorities to disprove misinformation and disinformation that has already been rampant among people to effectively terminate the spread. A real case history, from some years ago, concerns the rumours on Twitter about the presence of Ebola in Iowa, that started to worry people. Then, the Iowa state government had to release an official communication to ensure the population that no cases of Ebola were in the state.

2.3.2 ML @ work

In the following chapters ML techniques are often used to achieve the research objectives. Hereafter we provide a detailed explanation on how the experimental sessions were designed, the adopted validation strategies, the used methods and tools.

The type of ML experiments, involved in this research work, refers mainly to the supervised learning category, specifically, classification and regression tasks. The experimental sessions were planned to train, and subsequently compare the performance, predictive/decision models using the well-known algorithms in the literature.

To evaluate the models' performance a strategy called *cross-validation* has been adopted. This method divides the set of instances into a certain number (k) of disjointed subsets called *folds*. *k-1* folds are used for the training phase while the *k-th* fold is used for the model's test phase. Depending on how the instances in the *k-th* fold have been classified (correctly or not), the following information are derived: (for instance in binary classification task like bot detection): *true positive* (TP), i.e., the bot instances correctly recognized as bots; *true negative* (TN), i.e., the humans correctly recognized as humans; *false positive* (FP – type I error) i.e., the human accounts wrongly recognized as bots and *false negative* (FN – type II error) i.e., those bots accounts wrongly classified as humans. These four sets constitute the *confusion matrix* by which classification measures can be calculated. Briefly, TP and TN indicate the number of instances correctly classified, while FP and FN the instances wrongly classified. The most well-known and used metrics are:

accuracy (acc): $\frac{TP+FN}{TP+FP+FN+TN}$;

precision (P): $\frac{TP}{TP+FP}$;

recall (R): $\frac{TP}{TP+FN}$;

F1 [110]: $2 \times \frac{P \times R}{P+R}$;

A further important indicator of the model's effectiveness in classification, especially in case of unbalanced datasets, is the Receiver Operating Curve (ROC) [56]; more precisely, in the score called Area under ROC (AUC) whose value ranges from 0 (bad classification performance) to 1 (good classification performance). This process is repeated k times and the final result is given by the arithmetic mean calculated on the output scores obtained from the measurements on each k-th test fold.

In certain cases, the problem to be addressed is not the classification into categories but the prediction of a numerical value; in this case we refer to regression tasks. For this category there are specific learning algorithms, very different from those used for classification tasks, able to build a mathematical function/model to predict numerical values. Although the model validation methodology remains valid, the measures to evaluate the effectiveness of such predictors are not. In fact, it is not possible to use the measures listed above but others must be used. In these tasks, the most used measures are based on the quantification of the error between the values predicted by the model and the actual values of the dataset, the most known are: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

For both classification and regression tasks, once found the learning algorithm that produces the best results (e.g., accuracy and F1 in case of balanced datasets or AUC for unbalanced ones), it is possible to further improve the outcomes through a process called *hyper-parameter tuning*. Almost all learning algorithms, both in regression and classification tasks, have parameters that, if properly set, can lead to a considerable gain in model performance. In the case of neural networks based on multi-layers perceptrons, for example, the number of hidden-layers, the number of neurons per layer, the activation function within neurons and the number of epochs, are just some of the parameters that can be tuned. Each learning algorithm has specific parameters and only a knowledge of the algorithm can allow the experimenter to proceed for a successful tuning phase.

In our experiments we adopt the well-known ML framework called *Weka* [74]. In this framework we mainly used two items called: *explorer* and *experimenter*. The former concerns all the steps of KDD process described in the beginning of this section. It is possible to: handle data, train and test models (both classifiers and regressors), (hyper-parameter) tune the models, perform investigation on the extracted features and many other activities (like visualisation of models and performance). The latter allows to perform more complex experiments. In particular, it is possible to conduct comparative analysis among several learning algorithms (and/or datasets) to understand their performance and to see which models perform better than others. Roughly, the comparison between models is possible by setting up an algorithm (and consequently, the

related trained model) as *baseline* and, according to a specific (of aforementioned) measure, to observe if other models perform better than the *baseline*. To assess whether the difference, between the baseline and the other algorithms, is statistically significant the framework uses the *paired t-test*. It is also possible to perform two versions of the statistical test, i.e., *"corrected"* and *"uncorrected"*. The latter assumes that the samples are independent. However, due to the way cross validation works in the framework, this assumption is not always valid. Disregarding this, by using the uncorrected version, we can get *type I errors*. As opposite, the *corrected* t-test uses a *fudge* factor to counter the dependence between samples, resulting in more acceptable *type I errors*[6]. In general, it is strongly recommended to use the most reliable *corrected* version.

The way to set a baseline (to compare the trained models) mainly depends on the presence (or not) in the literature of a already defined baseline for the studied problem (e.g., the accuracy of the bot detectors). If possible, the baseline to be used is derived from the state of the art. Otherwise, a classic approach, used to evaluate a *good predictor*, has to be applied. This is the case when new problems are faced, and the scientific literature does not have a solid and comprehensive background. In such situations, pseudo-predictors are used to predict the mean value (for regression task) or the mode (for classification task), in Weka it is called *ZeroR*[7]. If the own trained models achieve performance statistically better than the baseline, the models are reputed good.

[6]https://weka.8497.n7.nabble.com/Paired-T-Tester-corrected-in-Experimenter-td21849.html
[7]ZeroR: https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/ZeroR.html

Chapter 3

Identification of Credulous Users on Twitter

3.1 Introduction

In this chapter we aim to acknowledge the existence of a class of users, namely *credulous*, and provide a technique to automatically rank them by inspecting the nature (bot or human) of their *followees*. The conducted experiments and related results will contribute to give an answer to our first research question. It is worth noticing that the content reported in this chapter have been published in [11].

The effectiveness of bots in influencing public opinion is confirmed in the literature [57, 97, 98] and the will to fight these malicious entities has stimulated a vast body of work on bots recognition in social media [29, 68, 155, 160]. Moreover, it has been observed that the majority of genuine users normally do not check the reliability of articles they read and/or share from OSM [51]. Depending on the activities of their contacts, these users may well end up actively contributing, although unknowingly, to spreading (unchecked) potentially harmful content.

As explained in Section 2.1.3, in Twitter there are two types of social relationships: being followed by someone (i.e., *followers*) or following someone (i.e., *followees*). As far as users' influence issue is concerned, it is not useful to consider *followers* because being followed by other accounts cannot be considered as an active action, unlike following someone. For this reason, taking into account the *followees* relationship makes more sense, since it expresses the interest of a user to keep up with content's updates of other ones. In fact, when a *followee* publishes something, this content is immediately displayed on the dashboard of those accounts following it (its followers).

Starting from this chapter, we deliberately draw attention to human operated accounts

in OSM, precisely to those users particularly exposed to the malicious activities planned by bot networks, with a higher risk to become potential consumers of targeted disinformation. The key feature of such human users is to own and follow an unreliable network of social contacts, more specifically, regarding their *followees*. A user following a great amount of bots (i.e., being a *bot-followees*) runs more the risk to see content published by bots on their own dashboard, that could harmful in case of malicious/deceptive content. By abuse of language we refer to such category of human-operated accounts as *credulous* users [11].

3.2 Proposed methodology

For the identification of credulous users on Twitter we designed an approach composed of two processes.

The first process aims to produce a refined decision model for bot detection which will be applied in the second process for the classification of human users' followees. The refined decision model is built upon an existing bot detector along with its dataset of Twitter users [166], where to each user is associated a label indicating whether it is a genuine user (in other words, a human-operated account) or a bot. During this process, data crawling from Twitter is performed to update the initial dataset. Such information is then converted into a set of representative user features. Multiple subsets of these features are then used to experiment with different machine learning algorithms [52] and maximise prediction *accuracy*. Finally, the best combination (of features and algorithm) is selected to obtain an improved decision model for bot detection. Section 3.2.1 describes this process in detail.

The second process deals with our main task of identifying credulous users. We start from the updated dataset, considering the users previously labeled as humans only, and extend it with additional information about the followees of those users. We then exploit the revised decision model (obtained by the first process) to label each followed account as a bot or a human. We also introduce a set of rules to determine whether a genuine user is a credulous one. This process is described in detail in Section 3.2.2.

3.2.1 Revisited bot detection

Our study starts by considering a publicly-available supervised dataset of Twitter users along with a bot detector trained on it [166] (see Figure 3.1).

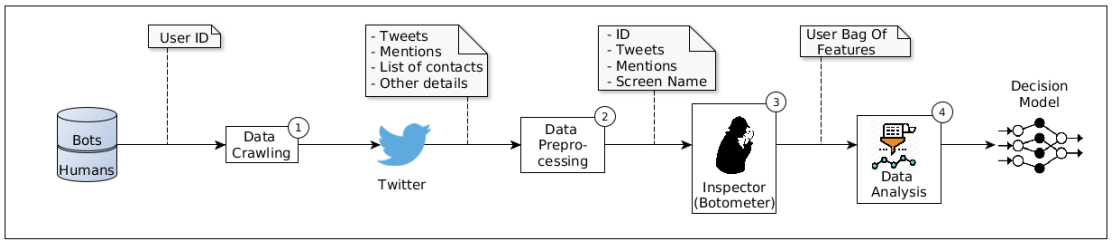


FIGURE 3.1: Revised Bot Detection.

Bot detectors have the tendency to gradually become obsolete and to loose precision, because bots evolve continuously [114], and existing datasets degenerate as time goes by, for instance due to suspended accounts. To partially overcome this issue, we decided to derive an improved decision model after updating the initial dataset. We run different machine learning algorithms on different sets of features, in order to compare their prediction accuracy. We eventually adopt as our refined decision model the best performing alternative among the considered ones on the basis of their *accuracy*.

Due to Twitter's policy restrictions[1], the initial dataset we relied on only contains the user IDs and the associated labels, indicating whether a given account corresponds to a human user or a bot. The IDs represent Twitter accounts, which we refer to as *basic users*. We implement *data crawling* on top of the Twitter API[2] to retrieve further information on those basic users (see ① in Figure 3.1).

For each basic user, our crawler fetches the following data :

- the *tweets*: the content in form of text, photos, etc. published by the user on his or her main page);
- the *mentions*: the tweets not published by the user, but where the user has been tagged by other users;
- the *list of contacts*: the list of the IDs of the users involved in any social relations with the considered user, i.e., *followers* and *followees* of a given user;
- *other details*: the *screen name*, the *description*, the *status count*, and other public information about a Twitter account.

During this step we also filter out from the initial dataset all the entries that are no longer valid, such as suspended or deleted user accounts.

After updating the initial dataset, the *data preprocessing* step (see ② in Figure 3.1) transforms the user data into a suitable format for querying the *inspector*. Requests to the inspector include the following user's data: *ID*, *tweets*, *mentions*, and *screen name*.

[1] **Twitter Developer Policy**: https://goo.gl/BiAG16
[2] **Twitter API**: https://goo.gl/2FXfi5

The inspector returns as output the probability of the user being a bot, along with a *bag of features*, i.e., a representative set of feature-value pairs (see ③ in Figure 3.1). For this task we rely on the Botometer web service[3].

To obtain our revisited bot detector we perform *data analysis* (see ④ in Figure 3.1). We thus compare the prediction accuracy in human-bot classification of different machine learning algorithms on multiple subsets of the features. The combination (of features and algorithm) showing the highest accuracy becomes our revised bot detector, i.e., the *decision model* of Figure 3.1.

To determine the probability of a Twitter account being a bot, the bot detector relies on six categories of features [166]:

1. *user-based*: the number of followees and followers, the number of tweets produced by the users, profile description and settings;
2. *friends*: the used language, local time, popularity, etc., extracted from followers-followees (i.e., retweeting, mentioning, being retweeted, and being mentioned);
3. *network*: the different types of communication (i.e., retweet, mention, and hashtag) weighted considering the frequency of interactions or co-occurrences;
4. *temporal*: the user activity (e.g., production of tweets) over different time intervals;
5. *content*: the natural language used and the length and the entropy of the text;
6. *sentiment*: the attitude or mood of a conversation, e.g., arousal, valence, and dominance scores.

At the time these experiments were carried out (in 2018), Botometer was at its first version. In that version Botometer produced as output only two *scores*: the so-called *english score* (ES) that relies on the six categories above and the *universal score* (US) that ignores sentiment and content features, being them English-specific.

We performed four different evaluations, see Table 3.1.

Features	**Algorithms**			
	C4.5+	*RF*	*R*	*NN*
ES	78.00	77.96	78.78	79.70
US	77.60	77.05	77.41	78.05
ES+US	78.23	73.02	78.65	79.83
Bag of features	81.16	81.71	81.30	**82.26**

TABLE 3.1: Percentage of correct prediction (*accuracy*) - bot detection.

[3]**Botometer web interface**: https://goo.gl/uyhG5c

The first three rows of the table refer to the outcome of the experiments based on the two scores (ES and US) separately and on their combination (*ES+US*). The fourth, more effective, experiment considers not only the values assigned by Botometer to the above six categories but also the numbers of tweets and mentions as further features. These eight features constitute our *bag of features*. The columns of the table list the considered algorithms: C4.5 [138], based on decision trees, random forests (RF) [23], RIPPER (R) [37], and neural networks (NN) based on the multilayer perceptron model with back propagation [70]. These algorithms are all well-known in the literature [44, 48, 166, 167]; moreover, the random forest algorithm has proven to be a rather accurate classifier for bot detection [166]. For the tests, we used the implementations available in the Weka tool-suite [52]. The values in the table represent the achieved prediction accuracy (expressed as a percentage) of models validated by means of the 10-fold cross validation. The highest value of 82.26% (bold text in the table) is obtained by using neural networks trained with the selected bag of features.

For clarification purposes, we would like to point out that the revised bot detector is composed of the model trained using the bag of features and based on a neural network. Unlike Botometer (merely used to obtain our *bag of features*), which returns only a (numerical) score indicating the probability of being a bot or human, our bot detector is able to assigns a class to each account. For this reason we called the resulting bot detector as "revised". It is used in the next process to determine whether a Twitter account, more precisely a user followed by a basic user, is a bot or a genuine one.

3.2.2 Identification of credulous users

The process used for identifying credulous users is shown in Figure 3.2.

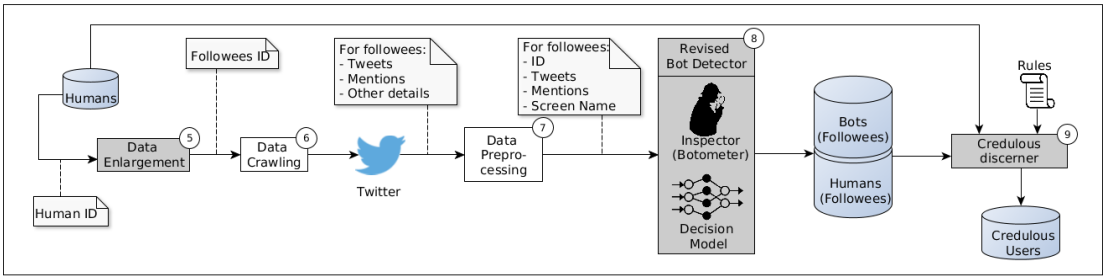


FIGURE 3.2: Identification of credulous users.

It starts by considering the human users of the same dataset [166] previously used in Section 3.2.1. This time, however, we are interested in analysing the followees (i.e., the users followed by a basic user, see Section 2.1.3) of these humans.

Let us denote by U the set of basic users in the initial dataset, by H_b the set of basic humans in U, and by B_b the set of basic bots in U. We additionally denote by $F(h)$ the set of users that are followed by a basic human (the human's *followees*) h, and by $F(H_b)$ the union set of the followees over all the basic humans in H_b. Therefore:

$$U = H_b \cup B_b \ , \ H_b \cap B_b = \emptyset$$
$$F(h) = \{f : h \in H_b \wedge f \ is \ followed \ by \ h\}$$
$$F(H_b) = \bigcup_{h \in H_b} F(h)$$

Due to the rate limits of the Twitter APIs and to prevent the overall number of (many potential) followees $|F(H_b)|$ from growing excessively, we only consider the subset H'_b of H_b whose users have at most 400 followees [12] on Twitter:

$$H'_b = \{h \in H_b : |F(h)| \leq 400\}.$$

This threshold has been primarily chosen not only to control computational cost reasons, i.e., to avoid a sharp increase in the number of followees to analyse. Further factors have been considered behind this choice, specifically: (i) the Twitter's limit in the number of accounts that can be followed per day[4,5] and (ii) the median number of accounts followed by the most active[6,7] Twitter users, i.e., 456 followees.

The *data enlargement* step (see ⑤ in Figure 3.2) consists in building, for this restricted set of basic human users, the overall list of followees to consider, i.e., $F(H'_b)$. The subsequent *data crawling* step (see ⑥ in Figure 3.2) is performed on the list of users $F(H'_b)$ produced above. This step is similar to the crawling step described in Section 3.2.1, except that here the list of contacts is not fetched. We perform a *data preprocessing* step that prepares the requests for our revised bot detector (see ⑦ in Figure 3.2). Specifically, information related to the followees of a genuine user (i.e., account details, tweets and mentions) is given as input to our revised bot detector module. For each followee, it first calculates the features via Botometer web-service (called *inspector*), then after adding the number of tweets and mentions through which Botometer derived the features, it builds the *bag of features*. The resulting instance is then processed by the decision model trained in the previous process in Section 3.2.1 (see ⑧ in Figure 3.2). This way, we compute a prediction p (i.e., 0 for humans, and 1 for bots) for each newly fetched user. In

[4]400 Twitter Statistics and Facts: https://tinyurl.com/y4flufbg
[5]About following on Twitter: https://tinyurl.com/y9ffq8lp
[6]Sizing Up Twitter Users (PEW research center): https://tinyurl.com/v3233b9
[7]With the term 'most active' Twitter users We refer to that 10% of Twitter users responsible for 80% of all tweets created by US users. See previous footnote of PEW research center.

the end, we are able to derive the set $BF(h)$ of *bot-followees* for every given human user h, and the overall set $B(H'_b)$ of *bots* for the whole set of basic humans users:

$$BF(h) = \{f \in F(h) : h \in H'_b \; and \; p(f) = 1\}$$
$$B(H'_b) = \bigcup_{h \in H'_b} BF(h)$$

In the following, we discuss the rules to rank the human users in H'_b. Each of these rules determine separate ranked lists of users that are eventually combined to build the set of credulous users.

The first rule (R_1) calculates, for a given basic human user, the ratio between the number of its bot-followees over the total number of its followees. Intuitively, this captures the observation that a user with a high number of bots in the list of followees is more likely to be influenced. This is expressed as follows:

$$R_1 = \frac{|BF(h)|}{|F(h)|}, \; h \in H'_b$$

The second rule (R_2) ranks users according to the normalized ratio between the number of bot-followees and the overall number of followees. Normalization is introduced to capture cases where humans have a high ratio of followed bots over their followees, but the actual number of followees is low in comparison to other users of the same dataset. The rule is:

$$R_2 = \overbrace{\frac{|BF(h)|}{|BF(h_{maxB})|}}^{bot \; normalization} * \overbrace{\frac{|F(h)|}{|F(h_{maxF}|)}}^{followees \; normalization}$$
$$h, h_{maxB}, h_{maxF} \in H'_b$$

where h_{maxB} represents the human with the highest number of bots among its followees, and h_{maxF} denotes the human with the highest number of followees.

The third rule (R_3) aims at giving relevance to the *seniority*, or *experience* of a user. Intuitively, more experienced users tend to follow new accounts by selecting them more carefully. For each basic user, the rule calculates the ratio between the value calculated by R_1 over the age (in months) of the account, denoted as age_m:

$$R_3 = \frac{R_1}{age_m(h)}, \; h \in H'_b$$

The fourth rule (R_4) considers the normalized relation between the number of bot-followees, followees, and age. The idea in this case is to capture the increased ability of younger accounts to effectively filter out more bots. Specifically:

$$R_4 = R_2 * \overbrace{\frac{age_m(h)}{|age_m(h_{maxA})|}}^{age\ normalization},\ h, h_{maxA} \in H'_b$$

where h_{maxA} represents the eldest human in H'_b.

On the basis of the above rules, we obtain four ranked lists of the users in H'_b in descending order. We additionally combine the four rules to understand which characteristics are more relevant.

We first study the usefulness of normalization, with two rules that embed normalized factors in their specification:

R_{13} considers the set of users selected by both R_1 and R_3. The idea is to prioritize users with a considerable amount of bots among their followees, but also take into account the percentage of bots with respect to their seniority. These two rules do not include normalized factors.

R_{24} considers the set of users selected by both R_2 and R_4. The rationale is to prioritize the normalization related to bots, number of followees, and age of a basic human.

We also investigate the usefulness of considering the age of the user accounts:

R_{12} considers the users selected by both R_1 and R_2. By doing so, we prioritize the information on the number of bots and followees, intentionally excluding the age.

R_{34} considers the set of users selected by both R_3 and R_4. Here we jointly consider the number of bots, the number of followees, and the age.

Finally, we combine all the four rules as follows:

R_{1234} considers the users jointly selected by all the rules combined together, so to observe the highest-ranked users with respect to all the provided rules.

The *credulous discerner* step (see ⑨ in Figure 3.2) applies all the rules reported above and produces ranked lists of users. Note that selecting the topmost users from these lists yields different sets of credulous users. We would like to point out that this process is not a decision model for classifying *credulous* users, it can be rather considered as a set of rules that contribute to understanding if a human is more exposed to bots than others. This represents an empirical investigation for building a preliminary dataset of *credulous* users. In the next section, we apply these rules to our dataset and discuss some experimental results.

3.3 Experimental results

In Section 3.2 we introduced and implemented a method to rank genuine users on the basis of their 'gullibility', so to isolate the most credulous ones from the rest. In the following we report our experimentation whose main purpose is to validate our gullibility ranking and its usefulness to single out credulous users. During our experiments, we noticed some other interesting findings that we discuss in the following.

We rank the genuine accounts according to the rules defined in Section 3.2 and generate different ranked lists for each rule. We thus obtain different sets of *potential* credulous users by selecting the topmost elements of these ranked lists. To quantify the usefulness of these sets, we define a measure of *efficacy* as the ratio between the number of detected bots over the total number of followees for the considered set of credulous users. We recall that these sets represent a first source of knowledge to further investigate the features of the user accounts, and single out credulous users.
Having applied the selection criteria from Section 3.2.2, we obtain from the dataset [166] 754 human users to be considered. This is our dataset D_1. Furthermore, in order to check that the obtained results are not dependent on the specific dataset, we build from D_1 two smaller datasets, D_2 and D_3, by randomly extracting (obviously, without re-injection) a half and a quarter of the elements of D_1, respectively.

We performed a preliminary investigation on these three datasets by measuring the efficacy for all the humans. The 754 human users in D_1 turned out to have about 126k followees, 17k of which were marked as bots, leading to an efficacy of 0.14; D_2 includes 377 humans, 65k followees, and 8k bots, for an efficacy of 0.13; D_3 has 188 humans, exposing 35k followees, and 4k bots, hence the efficacy is 0.13. An interesting observation is that all these values confirm that the claim about roughly 15% of Twitter users being bots [61, 166] holds also for the induced network of followees in the considered dataset. Another thing worth noticing is that our approach is expensive, as it required scanning about 126k user accounts to determine gullibility of 754 genuine users only.

Tables 3.2–3.5 report our experimental results. The datasets are reported in the leftmost side of the tables, where columns *size* and *id* report the number of users and their identifiers, respectively. Each table corresponds to a different group of experiments. The first one investigates the efficacy associated to the evaluation of the four rules of Section 3.2.2 in isolation and the related results are shown in Table 3.2.

To calculate efficacy, we introduced some cutoffs on the number of genuine users to be considered as (potential) credulous users. For example, for the D_1 dataset, we set three cutoff values to 200, 150, and 100 (shown in column *cred* of Table 3.2). By setting the cutoff to the topmost 200 users, the four rules yield an efficacy of 0.275, 0.197, 0.265, and

Datasets		**cred**	**Specific rules** (eff.)			
id	size		R_1	R_2	R_3	R_4
D_1	754	200	0.275	0.197	0.265	0.189
		150	0.303	0.217	0.289	0.204
		100	0.344	0.236	0.324	0.228
D_2	377	100	0.273	0.183	0.270	0.173
		75	0.305	0.199	0.291	0.183
		50	0.350	0.224	0.326	0.214
D_3	188	48	0.276	0.197	0.264	0.191
		36	0.308	0.223	0.293	0.213
		24	0.353	0.247	0.333	0.233

TABLE 3.2: Efficacy scores – Rules in isolation.

0.189, respectively. We remind the reader that these values represent the ratio between the amount of analysed users' bot-followees over the total number of their followees. This means that, for example, by applying rule R_1 with the largest cutoff, 27% of these users followees turn out to be bots.

We can observe an increasing trend in the efficacy values for smaller cutoffs (i.e., 150 and 100). For example, in Table 3.2 we can see that the efficacy values are 0.275, 0.303, and 0.344 when considering 200, 150, and 100 credulous, respectively. We also observe similar trends for all the other specific rules (i.e., R_2, R_3, R_4). Among all the efficacy values for the four specific rules, it is possible to see that the best values are obtained by considering rule R_1 with a cutoff value of 100. Using R_3 leads to similar values.

After considering the different rules separately, we study some possible combinations of them. We select the topmost users (again with cutoffs at 200, 150 and 100 elements) from the ranked lists separately produced by the specific rules, and then consider as credulous those users obtained by intersecting the lists involved in all the considered combinations.

Datasets		**Normalization**			
id	size	R_{13}		R_{24}	
		cred	eff.	cred	eff.
D_1	754	152	0.296	174	0.200
		114	0.308	125	0.218
		71	0.367	82	0.243
D_2	377	78	0.302	86	0.184
		59	0.324	59	0.202
		37	0.368	42	0.228
D_3	188	36	0.304	40	0.204
		31	0.315	31	0.221
		18	0.368	21	0.248

TABLE 3.3: Efficacy scores – Rules dataset independent *vs.* rules dataset dependent.

The values reported in Table 3.3 are concerned with the second group of experiments. Here the goal is to investigate the dataset-independence capability of our ranking rules. To this purpose, we compare the efficacy of the conjunction of rules R_1 and R_3 (denoted by R_{13} in Table 3.3) that do not use normalization and the conjunction of the remaining two rules R_2 and R_4 (denoted by R_{24} in Table 3.3) that instead rely on normalization. It is worth remarking that the rules with *normalization* are dataset dependent because some factors have been obtained from a specific user having (in our dataset) the highest number of bot-followees or followees or age (in months). We have that the best efficacy values are obtained in the absence of normalization. For example, considering the dataset named D_1, with the initial dataset of 200 potential credulous users, R_{13} spot out a set of 152 credulous users, with an efficacy of 0.296. With the same dataset, R_{24} instead selects 174 users as credulous. In this case, the achieved efficacy of 0.2 is sensibly lower than R_{13}. The value of efficacy still reflects the same trend observed before, i.e., more permissive (initial) cutoffs lead to lower efficacy.

Datasets		**Seniority**			
		R_{12}		R_{34}	
id	size	cred	eff.	cred	eff.
D_1	754	101	0.271	67	0.284
		66	0.303	42	0.304
		36	0.337	23	0.350
D_2	377	45	0.273	28	0.299
		31	0.303	18	0.315
		18	0.344	13	0.342
D_3	188	24	0.278	18	0.288
		18	0.309	13	0.315
		9	0.368	8	0.340

TABLE 3.4: Efficacy scores – Seniority relevance in rules.

In the third group of experiments, we investigate efficacy of the combination of rules R_1 and R_2 (denoted by R_{12} in Table 3.4) that do not consider seniority of users account w.r.t. the combination of rules R_3 and R_4 (denoted by R_{34} in Table 3.4) that, instead, consider the longevity (in months) of accounts. We notice that the best efficacy values are obtained by the rules that take seniority into account. For example, with the initial dataset (D_1) of 200 potentially credulous users, R_{12} builds a set of 101 credulous users with an efficacy of 0.271. With the same dataset, R_{34} singles out 67 credulous users, and the efficacy is 0.284, only slightly larger than that of R_{13}. The efficacy values for all entries of R_{12} and R_{34} are very similar, despite the considered number of credulous users that instead are much less for R_{34}. For example, considering the topmost 150 potentially credulous users from the ranked lists of rules in isolation, rule R_{12} selects 66 credulous users with an efficacy of 0.303 (2nd line in Table 3.4), whereas rule R_{34} only

considers 42 credulous users, but with basically the same efficacy. This suggests that considering seniority is useful for the dataset under analysis, but has a limited impact.

Datasets		**All rules**	
id	size	R_{1234}	
		cred	eff.
D_1	754	63	0.294
		37	0.320
		19	0.370
D_2	377	27	0.304
		17	0.323
		9	0.378
D_3	188	16	0.309
		13	0.315
		6	0.376

TABLE 3.5: Efficacy scores – Selected credulous users.

In the fourth group of experiments, we evaluate the combined efficacy of all the four rules; the related results are reported in Table 3.5. We selected the topmost 200, 150 and 100 potentially credulous users in ranked lists produced for the specific rules separately (from Table 3.2) and then considered their intersection (D_1 in Table 3.5). This cuts down the number of selected credulous users to 63, 37, and 19, respectively. Such a large reduction of the dataset size, due to the intersection, shows that each specific rule has the effect of classifying as credulous different genuine users. By observing the columns of efficacy values in the tables, we can see that by combining all the four rules we obtain the best results. Furthermore, the observation about the increase in efficacy for smaller cutoffs is still valid. In conclusion, the larger efficacy for smaller cutoffs, consistently observed over all our experiments, substantiates the validity of our proposed gullibility ranking.

3.4 Discussion

Being aware that in OSM human users' opinion can be biased by the malicious activities of bots, in this chapter we focused on defining a method for singling out a particular category of genuine (human) users. Specifically, we focused on those users that can be the potential subjects of targeted mis-/dis- information, calling them *credulous* users. The harmfulness of this kind of genuine users mainly rely on their gullibility about the nature (human or bot) of accounts they follow on OSM, leading them to own a considerable amount of bot-followees.

In this chapter, we have proposed an approach that identifies, via a 'gullibility' ranking, the most credulous users out of a set of human-operated Twitter accounts. The experimental results and the related evaluation procedure confirm the validity of our ranking mechanism both in terms of *efficacy* and shows that its not dataset dependent. The efficacy score can be assumed to be very close to what could be defined as a *gullibility* measure. We thus provide an answer to the **RQ1** below.

> **RQ1** – Among human Twitter users, which type of social relationship (e.g., following or being followed) is the most influential, and why? Does it make sense to assign a *gullibility* score to human users? Which user-related aspects should be taken into account in such a score? Does a clear separation between credulous and not credulous users exist? Or, simply, is one user more credulous than another?

ANSWER *– On Twitter the followees relationship, unlike followers, is the best expressive one (i.e., being influenced by other users) because it manifests the interest of an user to be updated about some other accounts, therefore to their content production.*
Since the high efficacy value, that indicates the bot-followees density among the credulous users' followees, we are confident about the usefulness of a gullibility score because it can interpreted as indication (to malicious OSM entities) about the convenience in performing mis-/dis- informative attacks on a user. Among the many aspects, useful for the identification of credulous users, we considered the ratio between the number of bot-followees and followees and the seniority; then, implementing them in the form of rules (i.e., R_1, R_2, R_3 and R_4).
Considering the rules in isolation (Table 3.2), it is not possible to make a clear distinction between credulous and not credulous users. The thresholds of potentially credulous users (namely, topmost users in each ranked list) have been chosen empirically, therefore they can only provide a rank where, according to the considered rule, one user can be defined more credulous than another. But, when the rules are jointly considered (by means of intersections among lists, e.g., Table 3.5), the difference between credulous and not credulous users is much clearer because of the large reduction in the number of users (credulous for all rules).

In the light of what emerged from these first investigations, it is advisable to highlight some details and limitations. First of all, it is worth to specify that being credulous is a non-observable state of the user, which the value of followed bots may help to infer. For this reason, we are cautious in stating that all the accounts identified by this method are credulous in the strictest sense (i.e., they follow bots only because of their naivety or incapability to distinguish them) due to the possibility that some accounts, created for bot monitoring purposes, may be identified as credulous.

However, it is worthwhile to remark that the robustness of this approach is strongly affected by the performance of the used bot detector in the task of classifying users' followees. In this context, restricting our investigation to only (human) users with at most 400 followees can be considered a further limitation. This threshold, even if set taking into account some (in our opinion deemed sensible) factors, can be reported as a parameter to be tuned. Indeed, it would be interesting to study how the variation of this threshold can influence the quality (in terms of efficacy) of the identified gullible users. Moreover, the application of this approach is very expensive both in terms of data (i.e., it requires scanning the potentially many users' followees), and computational time (i.e., mainly due to the Twitter API policy constraints). This limitation will be faced in the next Chapter 4. Furthermore, at this stage, we do not take into account to what extent these genuine users disseminate content and which accounts represent the sources of the propagated content. We will address this in Chapter 6.

Chapter 4

Automatic Detection of Credulous Twitter Users

4.1 Introduction

In the previous chapter we assessed the effectiveness and usefulness of identifying credulous users. However, we also remarked that the whole process is very expensive. In fact, checking all the accounts' followees implies high computational costs, and it may be unfeasible in case of a large amount of human-operated accounts.

Being aware of this limitation, in this chapter we explore the effectiveness of ML techniques to perform the same task (i.e., the identification of credulous users). Precisely, through the training of a set of decision models, we automate the recognition of credulous users. To this purpose, we investigate the *relevant* features of Twitter accounts that are able to distinguish credulous and not credulous users.

As mentioned in Section 2.3, in a typical supervised learning task, the starting point is the formation of a ground-truth. In this specific case, it is represented by human-operated accounts, labeled as credulous users or not. To the best of our knowledge, there are no publicly-available datasets of such a kind of Twitter users, and the one presented in Chapter 3 is no large enough. In this chapter, we overcome this issue by reapplying the approach introduced in Chapter 3, but starting from a much larger dataset.

The robustness of the credulous identification approach, used here to build a ground-truth of credulous users, depends on the capability of the employed bot detector in determining the human/bot nature of the humans' followees. Given the low performance of the bot detectors shown in the previous chapter (see Table 3.1), part of this chapter is aimed to train a more accurate bot detector.

The investigation conducted hereinafter focuses on the second research question. Some of the results shown in this chapter have been published in [12].

4.2 Approach

We set up three sequential processes: bot detection, credulous identification, and credulous classification. The first activity aims at training a binary classifier to recognise bots and human-operated accounts, with the goal to improve the performances shown in Section 3.2.1 (see Table 3.1). The second task focuses on the identification of credulous human-operated accounts, via the approach presented in Chapter 3, to build a ground-truth of credulous and not credulous users. The third step classifies human-operated accounts as credulous or not, by using the built ground-truth. Before detailing the three processes, we introduce the datasets used in this study.

4.2.1 Datasets

Besides the dataset used in the previous Chapter 3 (*VR17* below), we consider two further publicly available datasets[1], specifically:

CR15 presented in [44], it includes three smaller datasets. The first one has been collected for scientific purposes over a period of twelve days in December 2012, precisely from a research initiative named *@TheFakeProject*. It was created and advertised with the help of National newspapers. A specific Twitter account was created showing in its bio the motto "*Follow me only if you are NOT a fake*". Furthermore, to check if a follower is really a human, a verification phase has been conducted. In particular, an URL to a unique CAPTCHA has been sent to each follower as a direct message. Among the 574 accounts that started to follow it, 469 successfully pass a verification phase consisting in resolving a unique CAPTCHA. These 469 Twitter accounts were certified as human-operated. The second one was collected to carry on a sociological study focused on the strategic changes in the Italian political panorama between 2013-2015. 84,033 unique Twitter accounts used the hashtag *#elezioni2013* in their tweets: by performing a random sampling, 1,488 accounts were subject to a manual verification and labeled as genuine users. The third subset is composed of 833 fake accounts, bought from three different Twitter accounts online markets;

CR17 firstly appeared in [42]. Following a hybrid crowd-sensing approach [6], the authors randomly contacted Twitter users by asking simple questions in natural language.

[1] Bot Repository Datasets: https://goo.gl/87Kzcr

All the replies were manually verified and 3,474 Twitter accounts were certified as human-operated ones. Furthermore, the dataset contains 6,609 social spambots (e.g., spammers of job offers and advertising products on sale at Amazon);

VR17 firstly introduced in [166]. It consists of 2,573 Twitter accounts. A manual annotation was performed by inspecting the profile details and the produced content. Overall, 1,747 accounts were annotated as human-operated and 826 as bots.

From the merging of the three datasets, we obtain a unique labeled dataset (human-operated/bot) of 12,961 accounts - 7,165 bots and 5,796 humans. We use this dataset to train a bot detector, as shown in Section 4.2.2. To this end, we use the Java Twitter API[2], and for each account we collect: tweets (up to 3,200), mentions (up to 100), IDs of followees and followers (up to 5,000).

First, we need to detect the amount of bots which are followed by the 5,796 human-operated accounts. To do so, we need to crawl information (profile details, tweets and mentions) about the followees of such accounts. Due to the rate limits of the Twitter APIs, the huge amount of followees possibly belonging to the aforementioned human-operated accounts and faithfully acting like in Chapter 3, we consider only those accounts with a list of followees lower than or equal to 400 [11]. This leads to a dataset of 2,838 human-operated accounts, called *Humans2Consider* hereafter. By crawling the data related to their followees, we overall acquire information related to 406,810 Twitter accounts that represent the selected humans' followees.

4.2.2 Bot detection

A bot detection phase is required to discriminate bots and genuine accounts in the dataset of the 406,810 followees. The literature offers a plethora of successful approaches, based, e.g., on profile- [44, 143], network- [103, 172, 184], and posting-characteristics [29, 45, 67] of the accounts. However, also due to the capabilities of evolved spambots to evade detection [42, 43, 145], the performances of the diverse techniques degenerate over time [114]. Furthermore, some bot detectors are available online, but not fully usable due to restrictions in their terms of use (e.g., DeBot[3] [29]). To overcome these issues, we design and develop a supervised approach, which mixes features from popular scientific work and the ones introduced in Chapter 3 (see *bag of features* the Section 3.2.1).
In detail, we consider two sets of features listed in Table 4.1, where each feature is denoted by F and a number corresponding to its ID. The first one derives from the most updated version of Botometer [166, 183], precisely, the 3rd version [4]. In addition to the

[2]Twitter API: https://goo.gl/njcjr1
[3]DeBot api restriction: https://tinyurl.com/yapppq89
[4]Botometer: https://botometer.iuni.iu.edu/

original Botometer features [166] (i.e., the six categorical features listed in Section 3.2.1 – F20 - F25), we also include: the *CAP* scores[5] (namely, *Complete Automation Probability*, the novelty of the 3rd version – F28 and F29), the *Scores*[6] (F26 and F27), the number of tweets (F30) and mentions (F31); it is worth noting that the latter two features and the six categorical features constitute the *bag of features* used in Chapter 3. We call *Botometer+* this augmented set of features (F20-F31).

The second feature set is inherited from [44], where a classifier was designed to detect fake Twitter followers. We use almost all their *ClassA* features[7], except the one about '*duplicated profile picture*' (F13), (we have no means to verify whether the same profile picture is used twice in the whole Twitter-sphere). We call *ClassA-* this reduced set of features (F1-F19 w\o F13).

In the following, the conjunction of the two sets of features is referred as *ALL_features*.

Lbl	Feature Name	Description
		ClassA-'s features
F1	#friends/#followers^2	The ratio between the number of friends and the squared number of followers
F2	age (in months)	The number of months since the creation of the account
F3	#tweets	The number of tweets, retweets, replies and quotes of the account
F4	has a Name	True if a name is specified in the account's profile
F5	#friends	(Alias #followees): The number of accounts a user is following
F6	URL in profile	True if a URL is specified in the account's profile
F7	following rate	The number of followees over the sum of followees and followers
F8	default image after 2m	True if the account did not change the default image provided by Twitter in the account's profile after 2 months of its creation
F9	Belong to a list	True if the account is member of, at least, one list
F10	Profile has image	True if the account has an image in its profile
F11	#friends/#followers ≥ 50	True if the ratio between the number of friends and followers is greater than or equal 50
F12	'*bot*' in bio	True if there is a clear declaration of being a bot in the account's profile

[5]Complete Automation Probability scores (eng) and (uni): https://tinyurl.com/yxp3wqzh
[6]English and Universal scores: https://tinyurl.com/y2skbmqc
[7]ClassA features require only information available in the profile of the account [44].

F13	Duplicate profile pictures	True if the profile's image is the same of that of other accounts (never considered)
F14	2 x #followers $\geq$ #friends	True if twice the followers is greater than or equal the number of followees
F15	#friends/#followers $\simeq$ 100	True if an account is following a number of accounts that is about 100 order of magnitude the number of accounts that follows it
F16	profile has address	True if a location is specified in the account's profile
F17	no bio, no location, #friends $\geq$ 100	True if: the account has no description in the bio and location fields of its profile and the number of friends is greater than or equal 100
F18	has biography	True if the biography is specified in the account's profile
F19	#followers	The number of the account's followers
F20	Sentiment	ranging from 0 to 1, this score relates to the emotion conveyed by a piece of text [166]

Botometer+'s features

F21	Friend	ranging from 0 to 1, this score relates to the interconnectivity between users in terms of retweeting and mentioning between each other. It does not take into account Followee/follower information [166]
F22	User	ranging from 0 to 1, this score is calculated by considering user meta-data; relates [166]
F23	Content	ranging from 0 to 1, this score extracted by processing tweets by looking to their entropy and lenght; relates [166]
F24	Temporal	ranging from 0 to 1, this score relates on the users' activity, e.g.,time elapsing between tweets and their distribution [166]
F25	Net	ranging from 0 to 1, this score relates to the different type of communication linking users by means of mentions, retweets and hashtag realizing a net for each type [166]
F26	Score (eng)	ranging from 0 to 1, this score indicates if an account to be most human-like (0) or bot-like (1), by considering also text analysis (english language)
F27	Score (uni)	ranging from 0 to 1, this score indicate if an account to be most human-like (0) or bot-like (1), without considering also text analysis (english language
F28	CAP (eng)	ranging from 0 to 1, this score represents an updated version of Score (eng) feature

F29	CAP (uni)	ranging from 0 to 1, this score represents an updated version of Score (uni) feature
F30	#Tweets4WS	number of tweets sent to Botometer web service on which it calculates the features (F20-F29)
F31	#Mentions4WS	number of mentions sent to Botometer web service to calculate the features (F20-F29)

TABLE 4.1: Features list and description.

For our experimentation, we adopt nineteen well-known learning algorithms. We run them on the *Humans2Consider* instances, trying the three sets of features (i.e., *Botometer++*, *ClassA-* and *ALL_features*). The classification capabilities of the trained bot detectors is validated by means of 10-fold cross validation. Then, the classification performances are evaluated by considering the following metrics (presented in Section 2.3.2): *accuracy*, *precision*, *recall*, F1 [110], and Area Under the ROC Curve (*AUC*) [56]. The classifier showing the best *accuracy* score will pass through Hyper-Parameter tuning to single out the best parameters configuration (an algorithm-dependent procedure). This is to further improve the classifier's bot recognition skills. The tuned classifier will then used to label the humans' followees in *Humans2Consider* dataset (Section 4.2.1). All experimental results are presented and explained in Section 4.3.

4.2.3 Identification of credulous Twitter users

The identification of credulous users can be performed with multiple strategies, since there are various aspects that may contribute to spot those users more exposed to the malicious activities of bots. As anticipated in Section 2.2.1, there are not so many approaches focusing only on human-operated users, since most of the related work target the detection of bots, see, e.g., [46, 69]. However, it has been demonstrated [22, 147] that bots represent one of the primary means for diffusing and letting fake news become viral. Besides, on the very first phase of the news diffusion, human-operated accounts significantly contribute to the spreading of such news [51].

To discern whether a genuine user is a credulous one, we apply the approach described in Section 3.2 to the *Humans2Consider* dataset (2,838 human-operated accounts) and we get as output four lists of ranked human-operated users, one list for each rule.

Then, regarding each single rule (i.e., R_1, R_2, R_3, R_4), we select as 'potential' credulous users the topmost 753 from each ranked list (this selection is also referred as *cut*). The reason behind the *cut*'s cardinality comes from a simple proportional calculation. In the previous Chapter 3, with a dataset of 754 humans, we set the *cut* to the first 200 users for each ranked list. Proportionally, here, with a dataset of 2,838 humans, the *cut* have to be set at the 753rd user of each list. We consider as credulous users, all those

contained in the intersection between 'potential' credulous, on each ranked list (see *All rules* in Table 3.5). At the end of this procedure, we identify as *credulous* 316 users in *Humans2Consider*. This set represents the ground truth for the classification task presented in the next section.
It is worth to highlight that, to analyse 2,838 users, we needed 421k users' account information and 833 million of tweets; this corresponds to 3Tb of data in our DBMS.

4.2.4 Classification of credulous Twitter users

In this section we consider several decision models and learning algorithms, to find the most suitable to automatically classify a Twitter account as credulous or not. By using classifiers instead of the process described in Chapter 3, we can save further data gathering related to the humans' followees. Our ground truth, built in Section 4.2.3, involves 316 credulous and 2,522 not credulous users.
We perform the experiments as in Section 4.2.2. Specifically, we experiment the same learning algorithms, by training them with the instances of the *Humans2Consider* datasets and considering the same feature sets as those in Section 4.2.2 (i.e., *Botometer+*, *ClassA-* and *ALL_features*). The classification performance are evaluated by means of 10-fold cross-validation.
It is worth to stress that, unlike Section 4.2.2, where the dataset was almost balanced, now the learning algorithms take as input a very unbalanced ground truth, showing an unbalancing factor of 1:8 (namely, for 1 credulous instance there are 8 not credulous instances). In the literature, several strategies have been provided to deal with unbalanced datasets in machine learning tasks; the most popular are oversampling [30] and undersampling [104]. Roughly, oversampling fabricates new, artificial data, by taking into account the data distribution of the minority class. Well-known techniques for oversampling are: random (just duplicating instances), SMOTE [30] and ADASYN [76]. At first sight, this solution could seem suitable to overcome the drawback of the composition of the ground truth, but we decide not to apply this technique. The reasons are as follows:

1. few instances in the minority class (only 316 credulous users). This means we would have to produce more than 2,000 new instances;
2. the generated data do not always respect the data type semantics. For instance, in case of natural numbers (e.g., #tweets), real numbers could be produced instead;
3. features are somehow semantically linked between each other, e.g., the numbers of followees, or the tweets, could depend by the seniority of the account. By generating instances in such a way, it could be possible to have accounts with

thousands of tweets and/or followers in a few days: this is not realistic and it can bias the learning phase (generation of outliers);

Undersampling methods select, among the majority classes, the same number of instances of the minority classes. The most known techniques of undersampling are: random (just selecting instances), cluster [102], Tomek links [165] and ensemble learning [188].

In our case, the selection of only 316 not credulous instances over 2,522 would lead to a dangerous loss of information about the not credulous population. Therefore, inspired by the *under-sampling iteration* methodology introduced for strongly unbalanced datasets [96], we avoid to work with unbalanced datasets by adopting the following strategy. From the majority class (not credulous users), we randomly select a number of instances equal to the number of credulous ones (i.e., 316), without reinjection. This subset is then unified with the set of credulous users producing a balanced 'subdataset' hereinafter referred to as *fold*. Then, we repeat this process on previously unselected instances of not credulous set, until there are no more instances (see Figure 4.1). This way, we split the sets of not credulous users into smaller portions unifying each one with a copy of the credulous users sets to obtain a collection of balanced subdatasets.

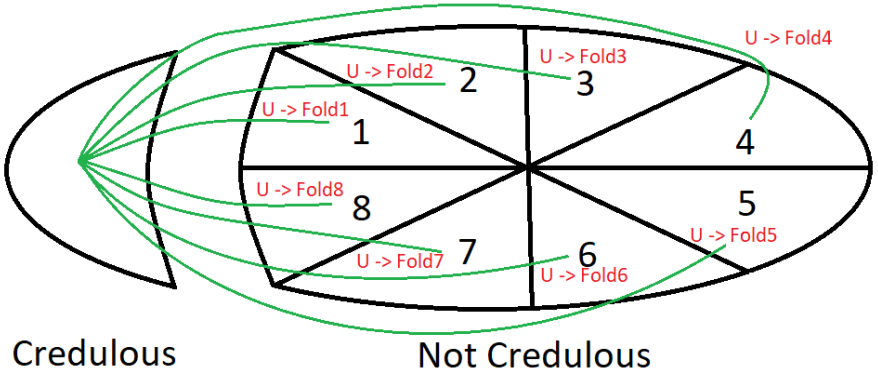


FIGURE 4.1: Adopted strategy to avoid an unbalanced set as ground truth. The credulous user set unified (U) with one of the not credulous users subsets produces (->) one fold.

For the sake of clarity, we detail the properties of *folds* as follows. Let us denote with n as the floor[8] of the ratio between the number instances in not Credulous users set (NC) over the number instances in Credulous users set (C). Let us define:

$$GroundTruth := C \cup NC;$$

and

$$NC_i := \{u \in NC : |NC_i| = |C|\} \qquad \text{for } i \in \mathbb{N}\, 1 \leq i \leq n$$

[8]The Floor function gives in output the integer part of a ratio.

and

$$NC_{n+1} = NC \setminus \bigcup_{i=1}^{n} NC_i$$

such that NC_i satisfy

$$\bigcup_{i=1}^{n+1} NC_i \quad = \quad NC \text{ and } NC_i \bigcap NC_j = \emptyset \text{ for } 1 \leq i < j \leq n+1.$$

Let us define

$$Fold_i \quad = \quad C \bigcup NC_i; \qquad \text{for } 1 \leq i \leq n+1.$$

It is trivial to obtain

$$\bigcup_{i=1}^{n+1} Fold_i = C \bigcup (\bigcup_{i=1}^{n+1} NC_i) = C \bigcup NC = GroundTruth$$

and

$$\bigcap_{i=1}^{n+1} Fold_i = C$$

Each learning algorithm is trained on each fold. To evaluate the classification performances on the complete dataset, and not just on individual folds, we compute the average of the values related to the folds composing each subdataset, for each considered evaluation metric.

4.3 Experimental results

All the experiments are performed with Weka [181], i.e., a tool providing the implementation of several machine learning algorithms. In the following, we present the main results obtained for bot detection and credulous classification.
The first column of Tables 4.2 and 4.3 shows the set of features considered for learning (i.e., *ALL_features*, *Botometer+* , *ClassA-*, see Section 4.2.2). The second column reports a subset of the adopted machine learning algorithms whose name is abbreviated according to the Weka's notation. The whole set of tested algorithms is as follows:

IBk: K-nearest neighbours [2], NB: Naive Bayes [85], SMO: Sequential Minimal Optimization [136], JRip: RIPPER [37], MLP: Multi-Layer Perceptron [127], RF: Random Forest [24], REP: Reduced-Error Pruning [137], 1R [79]

While the tables show the results of the most performing learning algorithms only, it is worth noting that 19 algorithms were used in the experimental phase. We refer the reader to the complete version of the experiments in the Appendix A.1.

The remaining columns in the tables report the values of the evaluation metrics (described in Section 2.3.2).

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	IBk	97.34	0.97	0.98	0.98	0.97
	NB	97.03	0.98	0.97	0.97	0.98
	SMO	98.04	0.98	0.98	0.98	0.98
	JRip	97.92	0.99	0.98	0.98	0.99
	RF	**98.33**	**0.99**	**0.98**	**0.98**	**1.00**
Botometer+	IBk	97.05	0.97	0.97	0.97	0.97
	NB	97.17	0.98	0.97	0.97	0.99
	SMO	97.64	0.98	0.98	0.98	0.98
	JRip	97.61	0.98	0.97	0.98	0.98
	RF	**97.97**	**0.98**	**0.98**	**0.98**	**1.00**
ClassA-	IBk	91.03	0.91	0.93	0.92	0.91
	NB	64.37	0.89	0.42	0.54	0.77
	MLP	85.01	0.89	0.84	0.86	0.91
	JRip	94.38	0.96	0.94	0.95	0.96
	RF	**95.84**	**0.98**	**0.95**	**0.96**	**0.99**

TABLE 4.2: Results for bot detection

Looking at the metric values, there is not a relevant difference in preferring a certain set of features over another. Albeit at a higher computational cost in features calculation, we prefer to adopt the features set that gives the highest performances. This choice is motivated by our need to apply the classifier to the humans' followees in the *Humans2Consider* dataset, with hundreds of thousands of accounts. Every single percentage point that we loose in accuracy translates into thousands of wrongly classified followees (in our case precisely 4k followees, since we have 406k followees – Section 4.2.1). Random Forest is the algorithm obtaining the best performances; with *ALL_features* (see the shaded line in Table 4.2), it achieves an accuracy = 98.33%, F1 = 0.98 and AUC = 1.00. After the hyper-parameter tuning phase, we obtain a slight improvement in accuracy of 98.41%.

In Table 4.3 are reported the performance results values obtained by running the learning algorithms on the *Humans2Consider* dataset, for the credulous classification task. It is worth noting that there are 316 Twitter accounts labelled as credulous users. At a first glance, we can see how *ALL_features* and *ClassA-* show to have good and quite similar classification performances, contrary to *Botometer+*. Both *ALL_features* and

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	IBk	89.69	0.74	0.73	0.90	0.96
	BN	80.26	0.91	0.89	0.76	0.91
	SMO	78.77	0.80	0.78	0.78	0.79
	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
	REP	93.07	0.99	0.88	0.93	0.94
Botometer+	IBk	65.03	0.61	0.60	0.63	0.70
	BN	61.02	0.67	0.62	0.49	0.69
	MLP	64.72	0.67	0.58	0.61	0.69
	JRip	66.42	0.67	0.67	0.66	0.67
	RF	**67.81**	**0.68**	**0.69**	**0.68**	**0.73**
ClassA-	IBk	92.59	0.74	0.73	0.92	0.97
	BN	82.77	0.98	0.88	0.79	0.93
	JRip	93.05	0.99	0.87	0.92	0.93
	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
	REP	93.09	0.98	0.88	0.93	0.95

TABLE 4.3: Results for credulous detection – 316 Credulous users.

ClassA- demonstrate their efficacy to discriminate credulous users. On the contrary, the *Botometer+*'s features properly work for bot detection tasks only. Going into deeper details, in Table 4.3 we can notice that the 1R algorithm obtains the best accuracy percentage (93.27% with $\sigma = 3.22$) and F1 (0.93).

As for the bot detector phase, Table 4.3 reports a part of the experiments. The full version is available in Table A.2, see Appendix A.2.1.

Finally, it is worth noting that the performances of the 1R algorithm are the same when considering *ALL_features* and *ClassA-*. This means that the algorithm selects *ClassA-*'s features only, the ones from *Botometer+* are useless in this case. This is a successful result, since we recall that *ClassA-* features refer to the profile of accounts only, and it is less expensive to calculate them.

4.3.1 Features analysis

This section extends the credulous classification analysis to assign each *ClassA-*'s feature an 'index of ability' to distinguish credulous from not credulous users.
Weka's tools assess the discriminatory importance of a feature in a set through the so called *attribute selection*. For the sake of reliability, we consider three attribute selector algorithms that evaluate the value (in terms of importance) of each attribute with

different methodologies: (i) *OneRAttributeEval*[9] uses the OneR classifier, (ii) *SymmetricalUncertAttributeEval*[10] measures the symmetric uncertainty with respect to the class and (iii) *InfoGainAttributeEval*[11] considers the information gain [90] against the class.

Rank	OneR	SymmetricalUncert	InfoGain
1	F1 (1.000)	F1 (1.000)	F1 (1.000)
2	F14 (0.977)	F14 (0.896)	F14 (0.894)
3	F19 (0.889)	F19 (0.509)	F19 (0.620)
4	F3 (0.768)	F5 (0.299)	F3 (0.323)
5	F5 (0.720)	F7 (0.235)	F5 (0.273)
6	F7 (0.712)	F3 (0.218)	F7 (0.255)

TABLE 4.4: Top most relevant *ClassA*-'s features (rank).

Table 4.4 shows the ranking of the first six most important features, according to the three evaluating algorithms. The remaining features have been estimated to impact with a lower relevance, in fact at least one of the evaluators estimated a value lower than 0.1, this happens for the seventh feature in the rank (i.e., $F9$) estimated as follows: 0.631 (*OneRAttributeEval*), 0.101 (*SymmetricalUncertAttributeEval*) and 0.085 (*InfoGainAttributeEval*). From Table 4.4, we can see that all the attribute evaluators confirm the relevance of the same features in the first six positions.

4.3.2 Further experiments

In the previous sections, we have proved the capability of ML techniques to successfully classify credulous users, thanks to the right selection of the discriminant features too. To further reinforce the latter findings, we decide to extend our experiments. The following experiments and related results are not included in [12].

What differentiates the following experiments from the previous ones is the number of users considered of being credulous in the *Human2Consider* dataset. In Section 3.2.2, starting from the four ranked lists (calculated by means of the four rules), 316 accounts were identified as credulous by intersecting the individuals in all the four lists. For each list, we consider as potential credulous users the first 753 users. The augmentation of this last value (potential credulous users) allows us to extend the number of credulous users. Since the number of the considered human-operated accounts remains unchanged (i.e., 2,838 in *Humans2Considered*), and what changes is the number of instances for each class (credulous vs not credulous users), we will refer to the next obtained ground-truths by calling them 'configuration'. The first cutoff is performed by selecting the first 946 users, i.e., considering one third of the accounts in *Humans2Consider* as potentially

[9] OneRAttributeEval: https://tinyurl.com/qtl3nox
[10] SymmetricalUncertAttributeEval: https://tinyurl.com/wcgccoz
[11] InfoGainAttributeEval: https://tinyurl.com/ve99qt8

credulous users. Then, by intersecting on the four ranked lists, we singled out 443 credulous users (the remaining 2,395 accounts are considered not credulous users). We call this configuration of the *Humans2Consider* dataset, *cut946*. Similarly, to obtain at least 500 instances of credulous users, we select the topmost 1,030 accounts as potential credulous users, producing a further configuration with 502 credulous users (and 2,336 not credulous users). We refer to this configuration as *cut1030*.
With these further experiments we aim at demonstrating that: (i) the introduction of 'fake' credulous users adversely affects the overall performances of the trained classifiers w.r.t. the results shown in Section 4.3 (see Table 4.3), (ii) *ClassA-* still remain the best set of features, and (iii) *Botometer+* features are still useless (given the very low accuracy values) for a credulous classification task.

Albeit in reduced form, *cut946* and *cut1030* suffer of unbalanced classes. Therefore, we apply the same strategy (i.e., iterative undersampling) to deal with this issue.

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	IBk	86.07	0.72	0.71	0.85	0.94
	BN	79.12	0.86	0.85	0.75	0.89
	SMO	78.48	0.80	0.74	0.77	0.78
	JRip	**89.92**	**0.97**	**0.82**	**0.88**	**0.91**
	REP	89.79	0.96	0.82	0.88	0.92
Botometer+	IBk	64.54	0.59	0.58	0.60	0.69
	BN	61.23	0.65	0.57	0.53	0.68
	MLP	64.84	0.66	0.56	0.60	0.69
	JRip	66.16	0.65	0.64	0.64	0.66
	RF	**66.36**	**0.66**	**0.64**	**0.65**	**0.71**
ClassA-	IBk	88.98	0.72	0.70	0.88	0.94
	BN	81.73	0.95	0.82	0.77	0.91
	RF	89.65	0.94	0.84	0.89	0.95
	JRip	**90.08**	**0.98**	**0.81**	**0.88**	**0.91**
	REP	89.91	0.94	0.82	0.88	0.92

TABLE 4.5: Results for credulous users detection – 443 Credulous users (*cut946*)

Tables 4.5 and 4.6 show part of the experimental results related to the two additional datasets with 443 and 502 credulous users, respectively. For the complete version of the whole experimental learning sessions, we invite the reader to read the appendix: Table A.3 for experiments on *cut946* (Appendix A.2.2) and Table A.4 for experiments on *cut1030* (Appendix A.2.3).

For the configuration with 443 credulous users (namely, *cut946* – Table 4.5), JRip performs better than the other algorithms, with an accuracy of almost 90% (89.92) using

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	IBk	85.23	0.71	0.70	0.84	0.93
	BN	78.24	0.84	0.84	0.75	0.89
	SMO	78.28	0.79	0.72	0.75	0.78
	JRip	**88.45**	**0.95**	**0.79**	**0.86**	**0.89**
	RF	88.31	0.92	0.81	0.86	0.94
Botometer+	IBk	66.31	0.57	0.57	0.61	0.70
	BN	60.87	0.65	0.58	0.55	0.68
	MLP	65.16	0.64	0.57	0.59	0.69
	JRip	**67.27**	**0.65**	**0.62**	**0.63**	**0.67**
	RF	66.87	0.65	0.62	0.63	0.71
ClassA-	IBk	87.25	0.71	0.70	0.85	0.93
	LAD	88.32	0.94	0.80	0.86	0.94
	RF	87.86	0.92	0.81	0.86	0.94
	JRip	**88.70**	**0.96**	**0.79**	**0.86**	**0.89**
	REP	88.39	0.94	0.80	0.86	0.92

TABLE 4.6: Results for credulous users detection – 502 Credulous users (*cut1030*)

ALL_ features and 90.08% using *ClassA-*. Once again, *ClassA-*'s features perform better than *Botometer+*'s ones (and slightly better than *ALL_features*). Compared to the dataset in Section 4.2.4, we can notice a slight degradation in the *accuracy* score, around 3%. Concerning the F1 and AUC, Random Forest (RF) obtains the best performances, slightly better than those corresponding to JRip; precisely 0.89 (F1) and 0.95 (AUC). Recalling that AUC is the best indicator to look at in case of unbalanced datasets (see Section 2.3.2), here we prefer to look to accuracy values because of the strategy adopted to overcome the issue of unbalanced datasets.

We conclude the section with comments on the performances obtained on the last configuration (*cut1030* – Table 4.6). This last configuration is the one with the larger number of credulous users (precisely 502). JRip obtains the best percentages of accuracy (88.45 with *ALL_ features*'s features and 88.70 with *ClassA-*'s features). As before, we confirm the usefulness of *ClassA-*'s features. Moreover, we assist to a further reduction of the best accuracy score w.r.t. the previous case. The highest F1 in *ALL_features* belongs to RF and JRip with an equal score of 0.86; in *ClassA-* F1 shows the same score with RF, LAD, JRip and REP. Concerning AUC, in *ALL_features* RF performs best (0.94) and in *ClassA-* the best score belongs to LAD and RF (0.94). As explained above, we consider as the best the algorithms showing the highest accuracy.

The results of these further experiments confirm our earlier expectation of a degradation about the classification performances related to the forced enlargement of credulous users. Furthermore, these results confirm that the *ClassA-*'s features are the most discriminatory ones. Despite the similar scores achieved by considering the union of features

sets (namely, *ALL_features*), the cheapness of *ClassA-*'s set is preferable, even if there was a slight degradation of the classification performances.

4.4 Discussion

The results shown in Tables 4.3-4.6 highlight the capability of our approach to automatically discriminate those Twitter users with a large number of bot-followees, namely credulous users.

We experimented two different feature sets (namely, *ClassA-* and *Botometer+*), plus their union set (called, *ALL_features*). *ClassA-* and *Botometer+* features need a very different amount of information for the feature engineering phase. In fact, *Botometer+* requires, for each user, all its tweets and mentions; while *ClassA-* needs the data in the user's profile, so called user's metadata[12]. On the latter set of features, we got the best classification results. The *Botometer+* set resulted to be less suitable to accomplish the credulous classification task.

Note that none of the adopted features have been derived by processing information related to the followees of the human-operated accounts (e.g., their tweets, mentions and profile data): this would have implied a high cost in terms of data gathering, storing and processing. These data have been only used to re-apply the approach in Chapter 3 (in order to get a larger ground truth of credulous users). This meant to retrieve information for more than 421k user accounts and 833 millions of tweets. On the contrary, the credulous detector, trained with *ClassA-*'s features, requires to gather the profile information of 2,838 accounts only. We underline that the features useful to discriminate credulous accounts are features belonging to the account profile only.

Even if the design of a bot detector was not our primary target, but only a mean through which building the ground truth for training the credulous classifiers, we notice that, compared to the performances reported in [42, 183], our bot detector achieves very good classification performances. This strengthens the robustness of the ground truth obtained in Section 4.2.3.

The findings in Sections 4.3 provide the answer to RQ2, shown below for convenience.

> **RQ2** – How effectively Machine Learning (ML) techniques can be in distinguishing credulous and non-credulous users? Is it possible to avoid in depth inspection of human users' social contacts in order to lighten the complexity of identifying credulous users? What is the loss in terms of accuracy when performing their identification? What are the features of Twitter accounts that

[12]Twitter User Object: https://tinyurl.com/y5s5kpuw

can facilitate this distinction? Are the features used for bot detection beneficial also for identifying credulous users?

ANSWER – *ML techniques have been proved to be effective in classifying credulous users. The classification results are very promising (e.g., in terms of accuracy and AUC values). This effectiveness pairs with great savings, both in terms of computational and time costs. The very effective features given as input to the classifier not only can be obtained by looking at the user's profile data only, but also they avoid the scan of the (potentially many) followees of the target account. This is a big advantage due to savings of computational costs (feature engineering phase) and time costs (data crawling). Although we tested several feature sets, we discovered that the most effective ones can be calculated from a user's profile information only (ClassA-). Through an evaluative analysis of those features we were able to single out the most determinant ones by checking several feature evaluation methods (see Table 4.4). The classification models have some limitations in terms of accuracy; but their achieved accuracy on average is 93.27% with σ of 3.22 (Table 4.3). We argue that such loss in accuracy can be considered more than acceptable. Lastly, we found that features designed for the bot detection task resulted to be useless for credulous users classification (Botometer+).*

However, although models based on RF achieve predictive accuracy values slightly lower than the best score (i.e., 93.27% with 1R model), it is worth to notice that, by considering the *ClassA-* features set, RF models get the best AUC scores (e.g., 0.97 in Table A.2, 0.95 in Table A.3, and 0.94 in Table A.4). An interesting aspect is that RFs perform well for both malevolent actors (e.g., bots as seen in Table 4.2) and victims (e.g., credulous users). The effectiveness of RFs in working well with profiles that exhibit a kind of stability is well-known in the literature [49, 146, 154]. This suggests the existence (for both roles) of some common features that are hard to disguise.

Chapter 5

Guessing the Number of Bot-followees

5.1 Introduction

In the previous chapter we mainly dealt with a classification task of supervised learning and we provided evidence of the effectiveness of Machine Learning (ML) techniques in automating the identification process of credulous users. In this chapter, starting from the definition of credulous users, we instead use ML techniques to determine, as precisely as possible, the quantity of bots a human-operated account is following, i.e., its *bot-followees*. Clearly, this task will be based on *regression analysis*.

Since credulous users are characterised by the fact that they follow many bots, it is not unreasonable to assume that the more bots a user is following, the more she/he is likely to read their misleading content, hence the more she/he is exposed to their potentially malicious activities. Guessing of the number of humans' bot-followees can be see as an extension to the credulous identification task and also as its generalisation since it aims to look for potential credulous users (see Section 4.2.3).
By using ML techniques, we aim at defining a regression model for predicting, the percentage of *bot-followees* of a human-operated account.
Having to deal with a problem that is conceptually different from a classification task, some of the learning algorithms (along with their performance measures), that will be used to train regression models, differ from the ones used in Chapter 4. What remains unchanged are the feature sets (namely, *Botometer+*, *ClassA-* and their union '*ALL_features*') representing the dataset's entries, and the dataset itself (in the previous chapter those were called *Humans2Consider* – Section 4.2.3). Changing the type of supervised learning task also changes the information associated with the dataset entries.

In fact, there is no longer useful now the information associated with being credulous or not (categorical class value), but rather, the percentage of bot-followees of each human-operated account (over its total followees).

We experiment on two "versions" of the *Humans2Consider* dataset. Firstly by considering the credulous users only, precisely the set with 316 instances, henceforth called *credulous-only* (see Section 4.2.3), and then on the whole set of human-operated accounts (with 2,838 instances, named *all_humans*). Despite *all_humans* contains all the entries in *Humans2Consider*, we use the latter name to refer at a simple set of human-operated Twitter accounts (just IDs, without any other information), and to the former name as the same set but with the added information concerning the bot-followees percentage (for each user). Although the main goal of this chapter is to determine the effectiveness of ML techniques in predicting the amount of *bot-followees* of a (human) user, we also find interesting to see whether knowing the information of being credulous (for a user) leads to more precise regression models.
Moreover, as done in Section 4.3.2, further investigations have been carried out by extending the experimental session running the learning algorithms also on to the other two groups of credulous users; i.e.,: *cut946* (with 443 credulous users) and *cut1030* (with 502 credulous users).

At the end of this chapter we provide our answer to the third research question. We want to let the reader know that the following experimental results and the related finding have been published in [8].

5.2 Experimental setup

This section explains the dataset, the features and the experimental design; specifying the metrics used to evaluate the performance of the trained models.

5.2.1 Dataset and features

In these experimental session, we adopt the dataset built in Chapter 4 and called *Humans2Consider* (see Section 4.2.1; we recall that it is composed of 2,838 IDs of human-operated accounts on Twitter [12] (publicly available [1]). Since now our task is to build regression models, that aims to forecast a numeric value instead of a categorical one, the information needed for each instance is not about being credulous but rather about the percentage of bots that each human user is following on Twitter (*bot-followees*). Recalling the data crawled from Twitter, here we do not rank human users (as done by means

[1] Dataset: https://tinyurl.com/y4o98c7l

of the rules in Section 3.2.2) but associate to each of them the respective percentage of bot-followees over the total number of contacts they are following (followees). This information can be obtained using the first rule (R_1) applied to the *Humans2Consider* dataset in Section 4.2.3. The difference between the ground-truth used in the previous Chapter 4 and the one experimented here (called *all_humans*[2]) is that the former reports binary values (about being a credulous user or not), whereas the latter includes Twitter accounts associated to the percentage of their bot-followees.

5.2.2 Experimental design

Our experiments start by setting up the data; precisely, by transforming the information related to the entries of our dataset accordingly to the three sets of features (namely *Botometer+, ClassA-* and their union set *ALL_features*).
Afterwards, to train regression models, 14 algorithms have been employed. To set up the experimental session and perform the experiments the machine learning framework called Weka [74] has been used. It is worth to notice that, by default, the results obtained by the *experimenter* in the weka framework are cross-validated. Accordingly to the notation adopted by the tool, the algorithms are: ZeroR[3] (used to obtain a baseline value against which to compare the performance of the other models), REPTree [137], LinearRegression [95], k-Nearest Neighbour (IBk) [2], LWL [5], AdditiveRegression [64], RegressionByDiscretization [62], M5Rules [139], DecisionStump [84], GaussianProcess [108], SMOreg [150], MultilayerPerceptron [127], MLPRegressor[4], RandomForest [23].
As anticipated in the introductory section, we experiment on two versions of our dataset. The first one includes credulous users only, i.e., those 316 singled out in Section 4.2.3. We refer to this ground-truth by the name *credulous-only*. Then, we experiment on all the human-operated accounts (*all_humans*).
Driven by the same motivation that led us to experiment on the set *credulous-only* and by simple curiosity, similarly to what we did in Section 4.3.2, we further extend our investigation to the other two sets of credulous users; precisely, by performing the same type of experiment on those users deemed credulous in *cut946*, with 443 entries, and in *cut1030*, with 502 entries (we used the same names in Section 4.3.2).

Results evaluation To the best of our knowledge, few papers address this problem, and even fewer approach it as a regression task [149, 169]. Because of this, no well-defined baseline is available from the literature to assess our results. To deal with this issue, we adopt the classic approach used to evaluate effectiveness of a predictive model [117], and

[2] *Ground-truth* (all_humans) : https://tinyurl.com/tcjjmbu
[3] ZeroR weka: https://tinyurl.com/y4hdhp54
[4] MLPRegressor weka: https://tinyurl.com/y5krc6d2

described in the last paragraph of Section 2.3.2. Applying that approach to the current context, we can compare the performance of the trained regression models (obtained from the aforementioned algorithms) with the score related to a *pseudo*-predictor obtained via ZeroR[5] method. This allows to predict the mean for a numeric class calculated on values into the ground-truth. Two metrics that are widely used for regression tasks[6] [117] have been considered to evaluate models' performance, namely Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE measures the average of errors in a set of numerical predictions, between the real values and the predicted ones ($MAE = \frac{1}{n}\sum_{j=1}^{n}|y_{real_j} - y_{pred_j}|$). RMSE measures the error too, and stresses more the prediction error by raising the square of the difference between the real values and the predicted ones ($RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_{real_j} - y_{pred_j})^2}$).

By using the *experimental-result analyser*, embedded in Weka [52], we performed statistical test (paired T-Test [82] with $\alpha = 0.05$) to determine which algorithm performs better than the baseline, relatively to each set of features.

5.3 Experimental results

We organise the experimental results in two sub-sections each centered on a specific dataset version. Section 5.3.1 reports the results about the experiments performed on the set with the 316 credulous users only (*credulous-only*). Section 5.3.2 shows the outcomes of the experiments with all the 2,838 human-operated accounts (*all_humans*).
Tables 5.1-5.4 have the same structure, but differ for the considered evaluation metric and the ground-truth version under investigation. The first column lists the algorithms mentioned in Section 5.2.2, while the remaining columns show the scores (related to the evaluation metric considered in the table) obtained when dataset's instances are represented according to a specific feature set; namely *Botometer+*, *ClassA-* and *ALL_features*. The first row of the table contains the baseline obtained through the *ZeroR* method. The star symbol, associated to some tables' entries, indicates that the score is significantly lower than the baseline, according to the paired t-test performed by Weka (see results evaluation in Section 5.2.2). In In all tables, the lowest score is reported in bold.

[5]ZeroR: https://tinyurl.com/wejh3vq
[6]https://tinyurl.com/yd9ljcmj

5.3.1 Credulous-only

Table 5.1 reports the scores related to RMSE. The baseline offered by *ZeroR* shows a score equal to 6.73%.

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*ALL_features*
→ ZeroR ← (baseline)	6.73	6.73	6.73
REPTree	6.92	6.86	6.86
LinearRegression	6.52 ↓	8.52	8.62
IBk	8.71	8.95	8.02
LWL	6.84	**6.10** ↓*	6.26 ↓
AdditiveRegression	6.79	6.30 ↓	6.20 ↓
RegressionByDiscretization	8.43	7.47	7.78
M5Rules	6.53 ↓	9.20	7.44
DecisionStump	6.90	6.15 ↓*	6.15 ↓*
GaussianProcesses	6.48 ↓	7.34	7.55
SMOreg	6.62 ↓	7.70	7.71
MultilayerPerceptron	10.79	11.97	11.82
MLPRegressor	7.59	7.50	6.86
RandomForest	6.60 ↓	6.15 ↓*	6.21↓

TABLE 5.1: RMSE scores – *credulous-only*

Note that in the column headed *Botometer+*, there are no starred values. Despite some of them are lower than the baseline (i.e., *LinearRegression*, *M5Rules*, *GaussianProcesses*, *SMOreg* and *RandomForest*), their values have not been considered significantly lower by Weka, in fact they are slightly better than the baseline. Among all values, the lowest one belongs to *GaussianProcesses* (6.48%) followed by *LinearRegression* (6.52%).
Differently from the *Botometer+* case, in the column headed *ClassA-* there are some starred values. Both *DecisionStump* and *RandomForest* achieve a RMSE score of 6.15%; but, the lowest one (6.10%) belongs to *LWL*, which also is the best RMSE score reported in Table 5.1.
The last column of Table 5.1 contains only one starred value which is also the column's lowest RMSE score (*DecisionStump*). Like for the preciding column, also here there are some values lower than the baseline (*AdditiveRegression* with 6.20% and *RandomForest* with 6.21%), but not enough to be starred.

Table 5.2 exposes the scores related to MAE. The indicated baseline is of 4.84%.
Looking at *Botometer+*'s features, the values lower than the baseline are: 4.83% (*LWL*, *AdditiveRegression*, *DecisionStump*), 4.68% (*LinearRegression*, *M5Rules*), 4.66% (*GaussianProcesses*) and 4.32% (*SMOreg*, starred). The latter is the lowest MAE.
When considering *ClassA-* features, the values lower than the baseline are: 4.78% (*LinearRegression*), 4.64% (*SMOreg*), 4.55% (*AdditiveRegression*), 4.54% (*RandomForest*)

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*ALL_features*
→ ZeroR ← (baseline)	4.84	4.84	4.84
REPTree	4.91	4.78 ↓	4.77 ↓
LinearRegression	4.68 ↓	5.18	5.19
IBk	5.88	6.17	5.44
LWL	4.83 ↓	4.36 ↓*	4.40 ↓*
AdditiveRegression	4.83 ↓	4.55 ↓	4.39 ↓*
RegressionByDiscretization	5.88	5.22	5.54
M5Rules	4.68 ↓	5.18	4.90
DecisionStump	4.83 ↓	4.36 ↓*	4.36 ↓*
GaussianProcesses	4.66 ↓	4.90	4.95
SMOreg	**4.32** ↓*	4.64 ↓	4.67 ↓
MultilayerPerceptron	6.91	8.17	7.92
MLPRegressor	5.16	5.19	4.78 ↓
RandomForest	4.86	4.54 ↓	4.44 ↓

TABLE 5.2: MAE scores – *credulous-only*

and 4.36% (*DecisionStump* and *LWL*, both starred).
When regression models are trained by using all features (column headed *ALL_features*), the values lower than the baseline are: 4.78% (*MLPRegressor*), 4.77% (*REPTree*), 4.67% (*SMOreg*), 4.44% (*RandomForest*), 4.40% (*LWL*, starred), 4.39% (*AdditiveRegression*, starred), 4.36% (*DecisionStump*, starred).

5.3.2 All_humans

Here we report the results related to the experiments performed on all human-operated accounts (2,838 instances).

Table 5.3 reports the values concerning RMSE, with a baseline value of 6.25%. In the second column (*Botometer+*), almost all the RMSE scores are lower than the baseline and starred, with the exception of *IBk* (7.73%), *RegressionByDiscretization* (6.32%) and *MultilayerPerceptron* (7.67%). With a score of 5.77%, *LinearRegression* has the lowest column's RMSE (and the second better one in Table 5.3). Concerning the third column (*ClassA-*), the situation is slightly worse. Despite nine values have better scores than the baseline, only three of them are significantly lower: 6.02% (*REPTree*, the lowest) and 6.06% (*DecisionStump* and *MLPRergessor* both). The forth column (*ALLfeatures*) shows a situation close to *Botometer+*'s case. In fact, with exception of four cases (i.e., *IBk*, *RegressionByDiscretization*, *MultilayerPerceptron* and *MLPRegressor*), all the other entries have significantly better values w.r.t. baseline (starred table's entries). The lowest RMSE of 5.72% is achieved by using *RandomForest* and it is the best score of Table 5.3.

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*ALL_features*
→ ZeroR ← (baseline)	6.25	6.25	6.25
REPTree	5.96 ↓*	6.02 ↓*	5.93 ↓*
LinearRegression	5.77 ↓*	6.14 ↓	5.80 ↓*
IBk	7.73	8.58	7.59
LWL	5.91 ↓*	6.08 ↓	5.99 ↓*
AdditiveRegression	5.84 ↓*	6.07 ↓	5.80 ↓*
RegressionByDiscretization	6.32	6.43	6.83
M5Rules	6.02 ↓*	6.16 ↓	5.84 ↓*
DecisionStump	5.96 ↓*	6.06 ↓*	6.02 ↓*
GaussianProcesses	5.79 ↓*	6.13 ↓	5.83 ↓*
SMOreg	5.91 ↓*	6.36	5.96 ↓*
MultilayerPerceptron	7.67	6.53	9.47
MLPRegressor	5.89 ↓*	6.06 ↓*	6.84
RandomForest	5.92 ↓*	6.09 ↓	**5.72** ↓*

TABLE 5.3: RMSE scores – *all_humans*

Finally, Table 5.4 presents the MAE outcome. The calculated baseline is 4.21%.

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*ALL_features*
→ ZeroR ← (baseline)	4.21	4.21	4.21
REPTree	3.95 ↓*	3.94 ↓*	3.87 ↓*
LinearRegression	3.84 ↓*	4.08 ↓	3.83 ↓*
IBk	5.07	5.43	4.95
LWL	3.97 ↓*	4.00 ↓*	3.98 ↓*
AdditiveRegression	3.89 ↓*	3.93 ↓*	3.76 ↓*
RegressionByDiscretization	4.16 ↓	4.24	4.36
M5Rules	3.91 ↓*	3.96 ↓*	3.82 ↓*
DecisionStump	4.06 ↓	3.99 ↓*	4.07 ↓
GaussianProcesses	3.87 ↓*	4.09 ↓	3.87 ↓*
SMOreg	3.67 ↓*	3.84 ↓*	**3.62** ↓*
MultilayerPerceptron	4.90	4.39	5.14
MLPRegressor	3.88 ↓*	3.93 ↓*	4.07 ↓
RandomForest	3.96 ↓*	3.96 ↓*	3.77 ↓*

TABLE 5.4: MAE scores – *all_humans*

At first sight, regardless of the feature sets, almost all the entries are lower than the baseline; and most of them are starred. When considering *Botometer+*'s features, the exceptions are: *IBk* (5.07%) and *MultilayerPerceptron* (4.90%); *RegressionByDiscretization* (4.16%) and *DecisionStump* (4.06%) are not lower enough to gain the star. For *Botometer+*, the lowest MAE is 3.67% (the second better value in the table) achieved by the model built with the *SMOreg* algorithm.

Similarly, when analysing the values of the column headed *ClassA-*, the values higher than the baseline are: 5.43% (*IBk*), 4.24% (*RegressionByDiscretization*) and 4.39%

(*MultilayerPerceptron*). The remaining values are significantly lower than the one in the *ZeroR*'s row but 4.08% (*LinearRegression*) and 4.09% (*GaussianProcesses*). Even in this case, the lowest score (3.84%) belongs to *SMOreg*.
Like in the previous cases, even when all features are taken into account, almost all values are lower, not only compared to the baseline, but also compared (line by line) to the values corresponding to the other two feature sets. The values overcoming the baseline are: 5.14% (*MultilayerPerceptron*), 4.95% (*IBk*) and 4.36% (*RegressionByDiscretization*). Apart for *MLPRegressor* and *DecisionStump* (both 4.07%), all the other entries have significantly better MAE values (starred). Once again, *SMOreg* outperforms the scores of the other algorithms for *ALL_features*, with a MAE of 3.62%; the lowest in Table 5.4.

5.3.3 Additional investigations

Although Tables 5.1–5.4 show that the information of being credulous is not only useless but also counterproductive, in this section we extend our experimentation to understand if this is due to randomness or to the set of credulous considered. To this end, like in Section 4.3.2, we further extend our experimental setting to the other two set of credulous users, i.e., *cut946* (443 users) and *cut1030* (502 users).

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*All_features*
→ ZeroR ← (baseline)	6.68	6.68	6.68
REPTree	6.95	6.84	6.37 ↓
LinearRegression	6.38 ↓	7.20	7.07
IBk	8.31	7.91	8.22
LWL	6.34 ↓	6.57 ↓	6.45 ↓
AdditiveRegression	6.26 ↓	6.04 ↓	**5.93** ↓
RegressionByDiscretization	7.84	7.34	8.47
M5Rules	6.81	7.17	8.38
DecisionStump	6.28 ↓	6.47 ↓	6.47 ↓
GaussianProcesses	6.37 ↓	7.01	6.73
SMOreg	6.55 ↓	7.00	6.65 ↓
MultilayerPerceptron	8.26	9.01	11.06
MLPRegressor	8.21	7.35	8.07
RandomForest	6.40 ↓	6.03 ↓	5.96 ↓

TABLE 5.5: RMSE scores – *cut946*

Table 5.5 reports the results related to the 443 credulous users (*cut946*) when RMSE is taken into account as evaluation measure. Although some of the values are lower than the baseline (e.g. *AdditiveRegression* with the lowest value of 5.93) these were not considered to be significantly lower (by the weka t-test). This is unlike Table 5.1, where, albeit few, there are some significantly lower values. Concerning MAE values, in

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*All_features*
→ ZeroR ← (baseline)	4.64	4.64	4.64
REPTree	4.73	4.50 ↓	4.40 ↓
LinearRegression	4.51 ↓	4.72	4.56 ↓
IBk	5.45	5.11	5.38
LWL	4.41 ↓	4.49 ↓	4.46 ↓
AdditiveRegression	4.43 ↓	4.27 ↓	4.17 ↓
RegressionByDiscretization	5.31	4.86	5.50
M5Rules	4.67	4.72	4.84
DecisionStump	4.40 ↓	4.51 ↓	4.51 ↓
GaussianProcesses	4.54 ↓	4.66	4.56 ↓
SMOreg	4.16 ↓*	4.28 ↓	**4.13** ↓*
MultilayerPerceptron	5.64	6.14	6.79
MLPRegressor	5.29	4.72	5.04
RandomForest	4.63 ↓	4.26 ↓	4.33 ↓

TABLE 5.6: MAE scores – *cut946*

Table 5.6, the situation is not much better. We got just two values significantly lower than the baseline and both of them were obtained with the same algorithm *SMOreg*, i.e. the algorithm by means of which we got the best MAE score in Table 5.2.

Algorithms	**Feature sets**		
	Botometer+	*ClassA-*	*All_features*
→ ZeroR ← (baseline)	6.57	6.57	6.57
REPTree	6.27 ↓	6.34 ↓	6.73
LinearRegression	6.27 ↓	6.49 ↓	6.14 ↓
IBk	7.91	8.10	8.22
LWL	6.14 ↓	6.54 ↓	6.43 ↓
AdditiveRegression	6.28 ↓	6.27 ↓	**5.98** ↓
RegressionByDiscretization	7.42	6.50 ↓	7.42
M5Rules	6.47 ↓	7.18	7.26
DecisionStump	6.16 ↓	6.62	6.66
GaussianProcesses	6.24 ↓	6.56 ↓	6.14 ↓
SMOreg	6.47 ↓	6.74	6.26 ↓
MultilayerPerceptron	9.01	7.76	9.91
MLPRegressor	8.11	7.17	8.72
RandomForest	6.21 ↓	6.05 ↓	5.99 ↓

TABLE 5.7: RMSE scores – *cut1030*

Table 5.7 reports the RMSE values obtained when the regression models are trained on the set with 502 credulous users (*cut1030*). Also in this case, we did not get values lower than the baseline despite some of them are under the baseline threshold. Finally, in Table 5.8, similarly to Table 5.6, the only values we got, that are significantly lower than baseline, are those related to the *SMOreg* algorithms.

Algorithms	Feature sets		
	Botometer+	*ClassA-*	*All_features*
→ ZeroR ← (baseline)	4.51	4.51	4.51
REPTree	4.41 ↓	4.46 ↓	4.45 ↓
LinearRegression	4.39 ↓	4.46 ↓	4.27 ↓
IBk	5.21	5.28	5.26
LWL	4.25 ↓	4.45 ↓	4.43 ↓
↓ AdditiveRegression	4.40 ↓	4.27 ↓	4.09 ↓
RegressionByDiscretization	4.99	4.55	5.09
M5Rules	4.45 ↓	4.67	4.51
DecisionStump	4.29 ↓	4.57	4.60
GaussianProcesses	4.40 ↓	4.41 ↓	4.27 ↓
SMOreg	4.08 ↓*	4.08 ↓*	**3.96** ↓*
MultilayerPerceptron	6.16	4.94	6.49
MLPRegressor	5.10	4.67	5.10
RandomForest	4.50 ↓	4.23 ↓	4.28 ↓

TABLE 5.8: MAE scores – *cut1030*

5.4 Discussion

In general we can state that better results are obtained when the full version of the dataset (*all_humans*) is considered rather than the version with *credulous* users only, in terms of both quantity, as number of models with performance better (hence, lower scores) than the baseline and quality, as the statistical significance of such scores. Such a situation can be explained by the fact that, performing a pre-classification phase (here in credulous and not credulous users), can somehow cut out some easy instances (useful to build an improved regressor) and worsen its overall performance.

However, focusing on baseline (1st line *ZeroR* of each tables) values of each metric, the highest ones (hence the worst) are observed in the version *credulous-only*. This is due to a higher distance between the real values (the right percentages of bot-followees) and the average value calculated on them. The fact that *all_humans*'s baselines have lower values than *credulous-only* indicates that the (*bot-followees*) percentages related to not-credulous users are more close to the average value than credulous users instances.

Considering RMSE, we got very few significant values lower than the corresponding baselines in both *all_humans* and *credulous-only*. Because of this and of the findings reported in [180], we prefer to assign more importance to MAE scores than to RMSE.

Considering the MAE metric, the models generated by the SMOreg algorithm are the most accurate ones, regardless of the considered dataset version. This concordance is limited exclusively to the algorithm, as these results have been obtained with two different feature sets: 4.32% by using *Botometer+* for *credulous-only* version (with 316 users) and 3.62% with *All_features* for the *All_humans* (which includes *Botometer+*'s features), *cut946* (4.13%) and *cut1030* (3.96%). While, considering the dataset with

all instances (all_hum), and looking to the MAE's scores obtained by SMOreg (with *Botometer+*'s features), we can see that the value is very similar (3,67%, the second-best result). This 0,05% loss (3,62 vs 3,67) can be overlooked due to the advantage of not having to calculate the ClassA- features (included in *ALL_features*). Therefore, at least as far as MAE metrics are concerned, a representation in *Botometer+* features combined with the use of the *SMOreg* algorithm can be considered the best choice.
In some cases, with the same algorithm, the score obtained by using *All_features* is identical (or very similar) to the one for *ClassA-* or *Botometer+*. This happens when the algorithm, during the construction and training phases of the predictive model, finds the features of a given set more effective than those of another one. Some examples are given by: *DecisionStump* in Tables 5.2 and 5.1, *GaussianProcesses* in Table 5.4 and *REPTree* in Table 5.1.
Further findings can be provided by studying, for each evaluation metric, the extent to which the choice of a feature set can affect the overall performance. Regardless of the cost to calculate a feature set, the experimental results do not show a great impact in preferring one feature set over another. Therefore, as far as the "MAE" is concerned, the previous assertion of preferring *Botometer+*'s features remains valid.
The outcomes of the above experiments allows us to provide an answer to our third research questions.

> **RQ3** – Is it possible to predict the number of bots a human user is following (*bot-followees*)? Are the features, used for credulous classification, useful also for this task? Which measures can be adopted to estimate the quality of such predictions in absence of well-defined benchmarks in the literature?

ANSWER – *From the fact that the results in Tables 5.1–5.8 are considered statistically significant by the adopted experimental ML framework (namely, Weka [74] – experimenter), we are quite convinced that the estimation of how many bot-followees a human-operated account is following (on Twitter) is far from being impossible.*
Although sometimes the features used for credulous classifications produce good results (see the starred values within column ClassA- in Tables 5.1–5.4,5.8), we should note that they are not our best achievements. In fact, the best performance scores are reached when ClassA-'s features (useful in credulous detection 4.3) and Botometer+'s features (designed for bot detection [49, 166]) have been jointly considered.
To the best of our knowledge, there are no well-defined benchmarks in literature with which to compare those obtained by us. For this reason and considering the claims in [180], we decide to adopt the MAE as main reference measure.

Although a priori classification did not lead to a more precise regression (in terms of MAE), it would be interesting to conduct further investigations in order to infer the

quantity of bots a genuine account is following; e.g., by considering the use of a joint approach between the classification and regression learning tasks (*joint learning*), successfully applied in age prediction [31].

Chapter 6

Credulous Users as Spreaders of Bot-originated Content

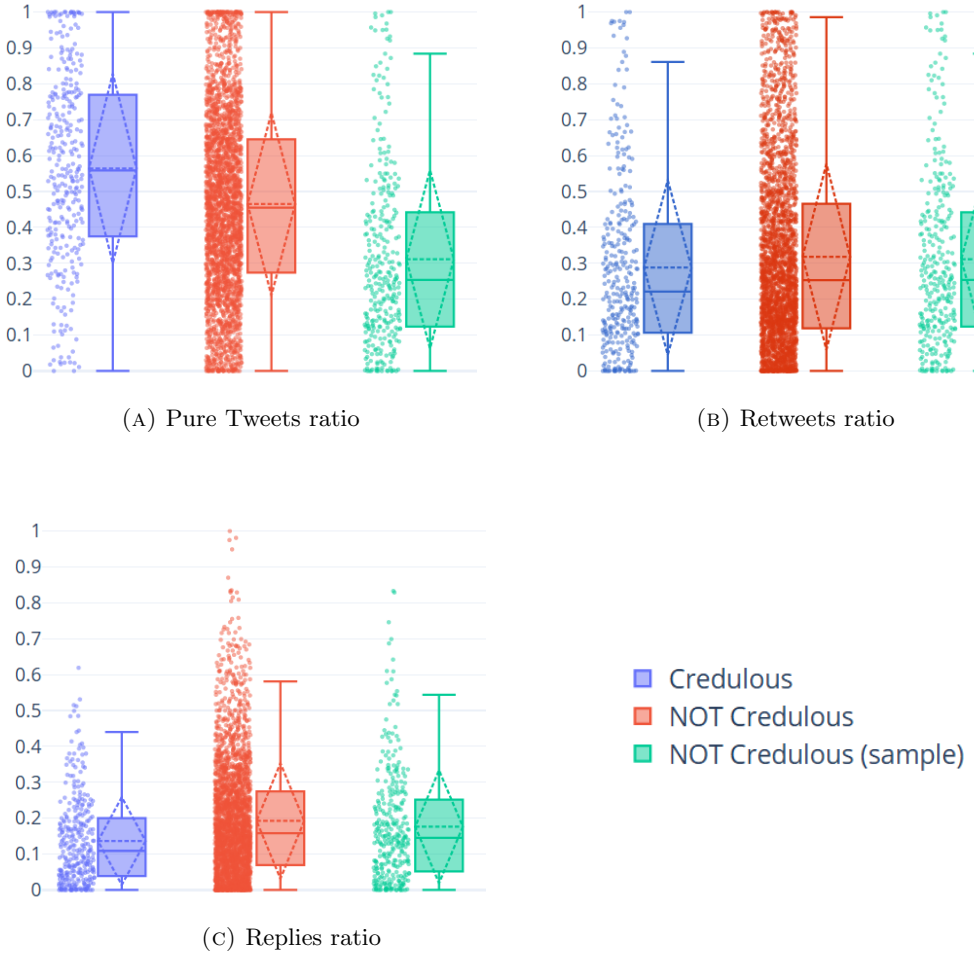
6.1 Introduction

In this chapter we start exploring the involvement of credulous users in supporting, even if unknowingly, bots' activities. We will compare the behaviour of credulous users with that of not credulous ones and provide evidence of a greater involvement of credulous users in the dissemination of content originated by bots.
Specifically, we will first perform a coarse-grained analysis to determine whether behavioural differences in terms of content production between C and NC users exist. More specifically, we will compare statistics and distribution, for each population, by focusing on their tweeting, replying and retweeting (considering quotes as retweets) behaviour/rate.
Then, we will conduct a fine-grained analysis by focusing on each tweet posted by C and NC users. Precisely, we will analyse all the tweets that the users in our dataset bounced via quotes, retweets and replies. For each kind of these tweets, we will trace back the Twitter author who first posted the original tweet and then check whether is is a bot or not. In this way, we will assess the level of engagement of C users in spreading potential malicious content (because bot-originated) with respect to NC users. We will use statistical tests to validate the significance of the behavioural differences between the population of C and NC users, if any.

Moreover, this analysis has the dual purpose of verifying the effectiveness of the credulous identification method (developed in Chapter 3 and used in Chapter 4) as well as the reliability and usefulness of the credulous classifier trained in Chapter 4.

In the following we present the experimental results and the related findings that rely on [9].

6.2 Behavioural analysis



(A) Pure Tweets ratio

(B) Retweets ratio

(C) Replies ratio

FIGURE 6.1: Activities of credulous users (*vs* not) – Distributions and stats.

In this section, we shed light on the activities of credulous accounts, in terms of tweets originated by users (hereinafter called *pure tweets*) (Figure 6.1a), *retweets* (Figure 6.1b), and *replies* (Figure 6.1c). It is worth noting that, in this first analysis, quoted tweets have been considered as retweets[1]. Results are shown in Figure 6.1. For each type of content, each subfigure reports statistics about the users' activities for the 316 C users (leftmost bar and points in each subfigure – in blue), the 2,522 NC users (bar and points in the middle in each subfigure – in red), and a random sample of NC users of the same number of C ones, 316 (rightmost bar and points in each subfigure – in green).

[1]On Twitter, a quoted tweet is a retweet with an extra text inserted by the retweeter.

Figure 6.1a reports the information related to *pure tweets*. Considering the overall amount of tweets, C users (blue points) produced, on average, the 56.44% of tweets (horizontal blue dashed line), with a standard deviation (dashed blue rhombus) of 26.4%. The totality of NC users (red points) feature an average tweets production that is lower than C users, precisely 46.49% (σ=25.45%). Looking at a sample of NC users (green points), we notice an even lower average (31.13%, σ=24.85%). The analysis of this first graph suggests that those accounts classified as credulous tweet more original content than the others.

Figure 6.1b reports the information related to retweets and quotes (w.r.t. the overall amount of tweets). In this case, the difference between C and NC users is less marked. C users (blue points) show a retweets-tweets ratio equal to 0.29(σ=0.24), while for NC users (red points) the ratio is 0.32 (σ=0.26).Very similar scores are obtained when considering the NC users' sample (green points) with average ratio =0.31(σ=0.25).

Similar findings have been obtained for replies, see Figure 6.1c. The replies-tweets ratio is equal to 0.14 (σ=0.124) for C users (blue points). The same ratio for the NC population (red points) is higher, with a value equal to 0.19 (σ=0.16). For the NC users' sample, we have a 0.18 (σ=0.16) ratio.

When considering retweets and replies it is more difficult, w.r.t. the case of pure tweets, to find differences between the two populations, as the relative average values are similar. Although for both retweets and replies the averages of C users are lower than those of NC users, we cannot say that the differences are as significant as in the case of pure tweets.
Therefore, the main result of this first (coarse-grained) analysis is that the posting behaviour, between C and NC users, is statistically distinguishable only in one case, the one in which tweets are produced by users themselves. On the contrary, no clear distinction can be made when retweets and replies are examined. Table 6.1 resumes in a numerical format the stats of Figure 6.1.

	Pure Tweets		**Retweets**		**Replies**	
	μ	σ	μ	σ	μ	σ
Credulous (C)	0.56	0.26	0.29	0.24	0.14	0.12
Not Credulous (NC)	0.46	0.25	0.32	0.26	0.19	0.16
NC (sample)	0.31	0.25	0.31	0.25	0.18	0.16

TABLE 6.1: Numerical overview of the stats pictorially reported in Figure 6.1.

For a deeper investigation, we will consider separately the quoted tweets and retweets, and then we will analyse the nature (human or bot) of the accounts that originated the tweets after the have been retweeted, quoted or commented by C and NC users.

Precisely, for each of the 2,838 human-operated accounts in our dataset, and for the three types of actions (i.e., quoted tweets, retweets and replies), we will calculate the percentage of content originated by bots. Considering, for example, the case of retweets, it is possible to retrieve the ID of the original tweet. Consequently, from the tweet ID, it is possible to retrieve the tweet author. We can then evaluate if that author is classified as bot or not. The same procedure is repeated for replies and quoted tweets.

For the bot classification phase, we adopt the bot detector presented in Section 4.2.2. The authors of the original tweets retweeted and quoted by our human-operated accounts, or to which they responded, are 1,22 million different users. Among them, 104,000 (8.5%) have been classified as bots.

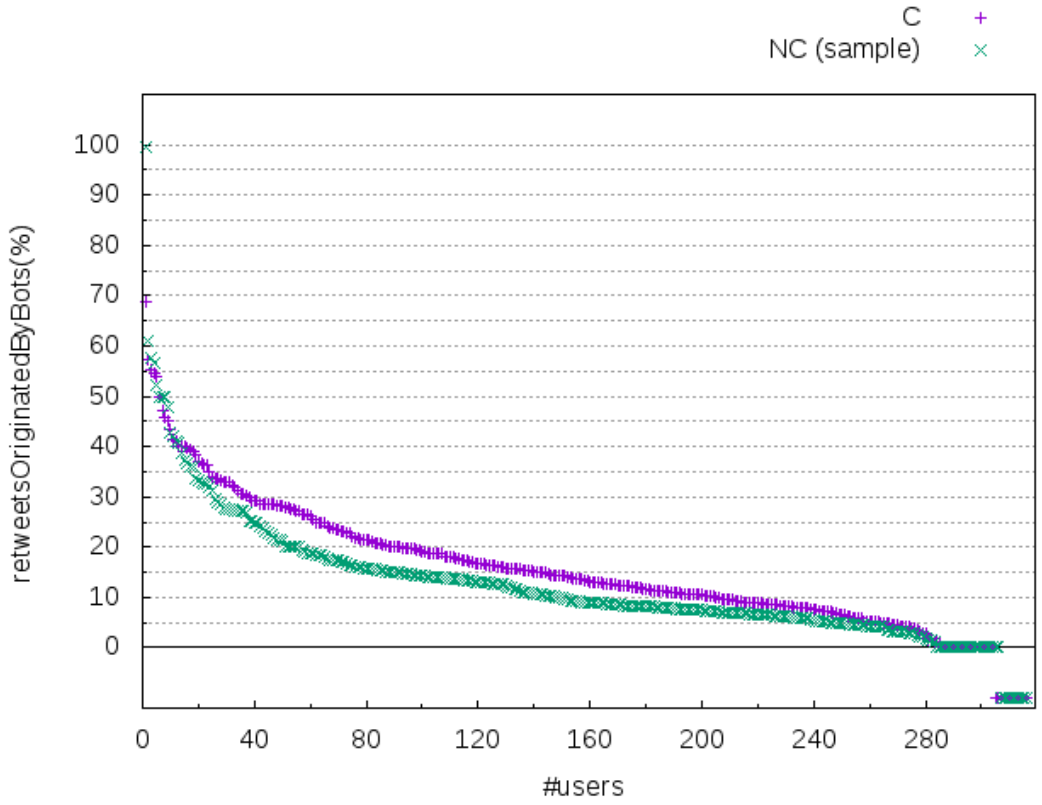
6.2.1 Retweets

Figure 6.2 gives two different views of the same phenomenon. In both the subfigures, C users are represented in purple, while NC users are in green.

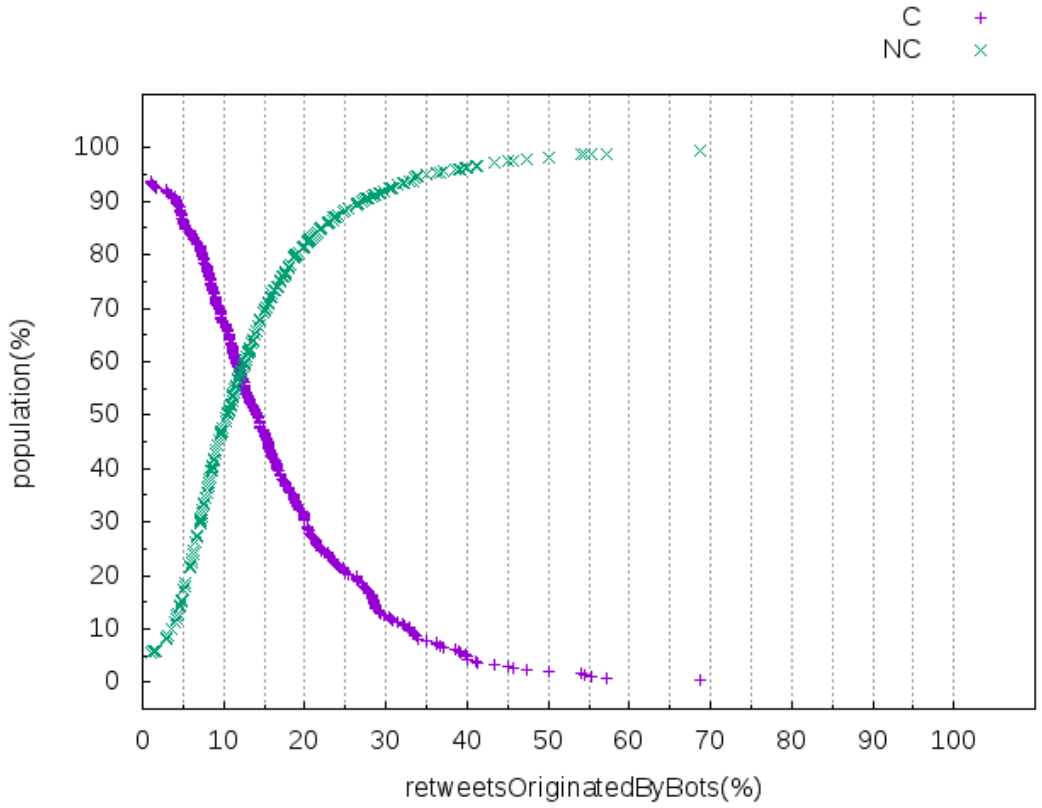
Figure 6.2a gives, on the y-axis, the percentage of retweets whose original tweets have been originated by bots[2]. On the x-axis, instead of reporting the Twitter ID of each users (which is a long string of numbers), we prefer to indicate them with consecutive numbers. Such choice is useful not only for the sake of readability but also to count the number of users with a percentage of byBot-retweets greater/lower than a certain threshold.

It is worth reminding that the original NC set is composed of 2,522 users; hence in the figure, for sake of a fair comparison, we consider just a (representative) sample of NC users, equal to the number of our C users (i.e., 316). To obtain a representative sample, we first built 20 samples of 316 NC users - each sample was obtained by randomly selecting instances from the original set, without re-injection. Then, for each sample, we computed the average and standard deviation on the percentage of byBot-retweets. Finally, we computed the Euclidean distance between the averages and standard deviations of the samples and we compared them to the ones calculated over the entire NC population. We identified as more representative the sample with the smallest distance. Looking at Figure 6.2a, we can notice that almost all the purple points (C users) are over the green ones (sample of NC users). The average percentage of byBot-retweets by C users is 16.45 ($\sigma = 11.84\%$), while the average percentage for NC users is lower, 13.41 (with $\sigma = 10.58\%$). The percentage of byBot-retweets have been calculated over the total amount of retweets. Some of the human-operated accounts in our dataset do not retweet at all. We call such accounts outliers. In Figure 6.2a, the outliers are shown

[2]Hereafter we will denote such retweets as 'byBot-retweets'.



(A) Percentage of 'byBots'-retweets posted by C and NC (sample) users.



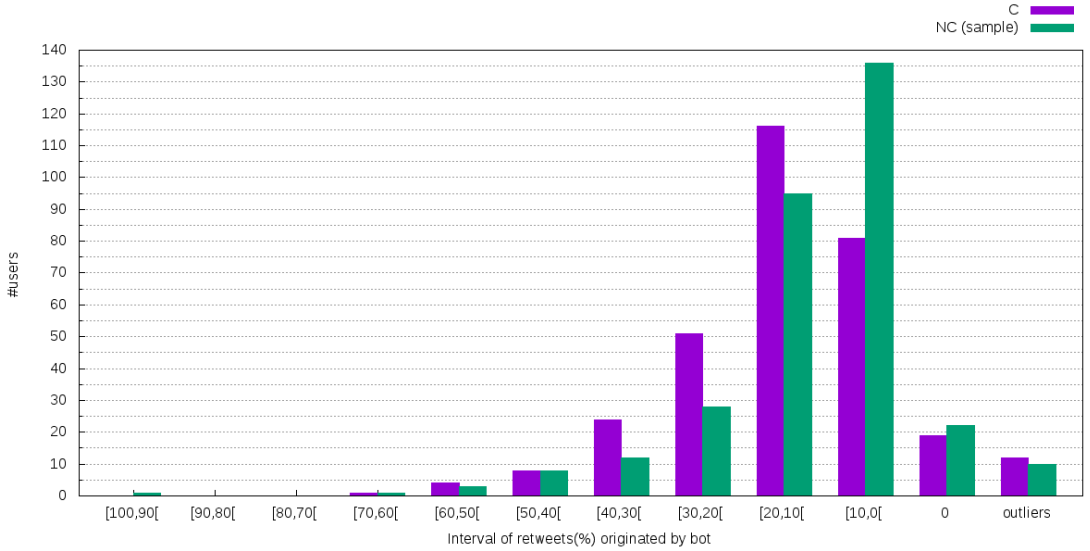
(B) % of populations w.r.t. the % of 'byBots'-retweets.

FIGURE 6.2: Comparative analysis between C and NC users w.r.t. 'byBots'-retweets.

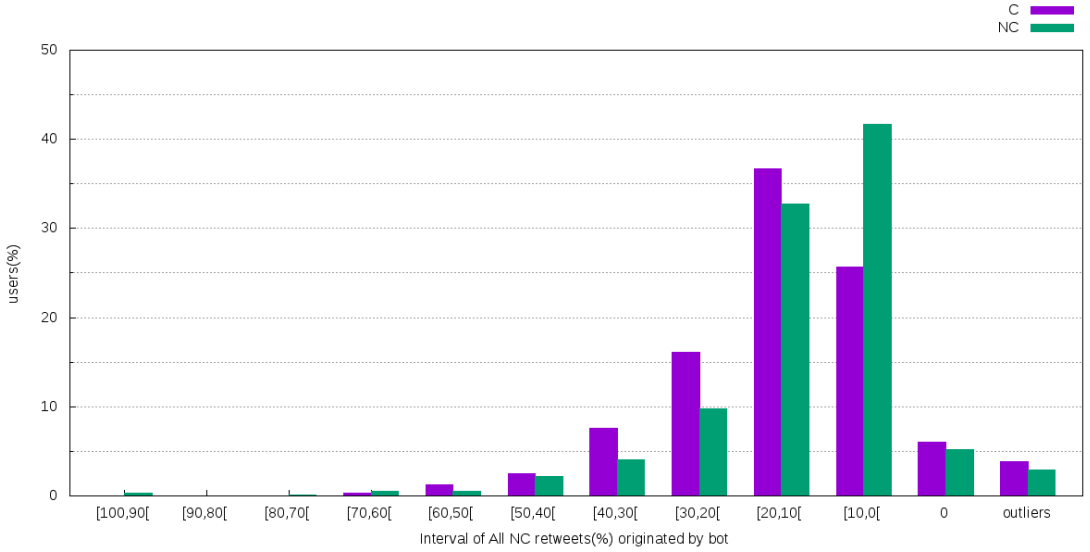
under the zero on the y-axis: 12 C users and 7 NC users are outliers. Moreover, the users lying exactly on the y-axis are those users who retweet only tweets originated by human-operated accounts.

Figure 6.2b compares the whole C and NC populations. The values on the x-axis are the same of those on y-axis in Figure 6.2a. Instead, on the y-axis, we report the percentage of the population having byBot-retweets (in percentage) *greater or equal to* (for C users – purple dots) or *lower than* (for NC users – green dots) the values on the x-axis. The aim of the graphs in Figure 6.2b is conveying a measure of *population coverage*, i.e., fixing the number of byBot-retweets as threshold, so that we know the percentage of C users whose byBot-retweets is larger or equal to the threshold, and the percentage of NC users which retweets is less than the threshold. In Figure 6.2b, the data related to NC users refer to all of them (2,522). It is important to stress that, both here and in the following figures of this same type, the abscissa point where the two data series intersect has no particular meaning and its values is not the one of *max* population coverage. The concept of max population coverage is strongly linked with a specific value on the x-axis, determined in the following way. For each percentage value on the x-axis, we sum the two corresponding percentage values on the y-axis, which correspond to the percentages of C and NC users population. The value on the x-axis where this sum assumes the absolute maximum (threshold) represents the point of max population coverage. This allows us to define the two biggest subsets of C and NC users where the former post more 'byBots' content than the latter. In this particular case, such a threshold point is The green and purple curves intersect at the abscissa 15.59. The 43.75% of C users has a percentage of byBot-retweets $\geq$ 15.59 (coordinates 15.59, 43.75 – purple dots). The 71.04% of NC users has a percentage of byBot-retweets $<$ 15.59 (coordinates 15.59, 70.04 – green dots).

Going further with the analysis, Figure 6.3 provides two aggregation perspectives, by grouping the C and NC users according to the number of their byBot-retweets. In Figure 6.3a, the x-axis reports the intervals (deciles) of byBot-retweets and the y-axis reports the number of users falling in each interval. Since the two sets (C and NC) have the same number of users (316), we prefer to report the real number of users, instead of the percentage. The sample of NC users is the same used for the results shown in Figure 6.2a. Figure 6.3b considers all the NC users. Since they are 2,522, we report the percentage (y-axis). When considering the whole population of NC users, we can notice that the comparison trend, for each decile,is preserved. This suggests that the subset of 316 NC users is a good representative of the NC population (Figure 6.3a). Finally, in both the subfigures of Figure 6.3, the users in the last group, i.e., the outliers, do not retweet any tweet. The users in the 0 group are those retweeting tweets originated by human-operated accounts only.



(A) Deciles of Figure 6.2a



(B) Deciles of C and all NC users

FIGURE 6.3: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-retweets

Findings From Figure 6.2, we can appreciate a difference in users' behaviour between C and NC users. On average, C users feature a higher percentage of retweets whose original tweets have been originated by bots. The difference between the standard deviation values for the two populations is negligible, indicating a behavioural similarity between C and NC users (Figure 6.2a). Since C and NC users are human-operated account, the similarity of standard deviations was expected. Both the subfigures in Figure 6.3a show a larger presence of C users in almost all the deciles; the only relevant difference is for the *[10,0[* group. In this group, there are more NC users than C users.

6.2.2 Replies

Figures 6.4 and 6.5 report the analysis related to the replies.

Figure 6.4a shows a quite clear difference between C and (a sample of) NC users. C users have an average percentage of replies to bot's tweets equal to 13.77 ($\sigma = 15.10\%$), while NC users show a mean's value of 10.28 ($\sigma = 11.68\%$). As for the retweets, the number of outliers is quite low (9 and 12 accounts for C and NC users, respectively). Figure 6.4b shows that the maximum percentage of covered population is achieved on a replies percentage value equal to 27.96 (x-axis). Specifically, the 11.40% of C users reply to bot's tweets more than the 91.56% of NC users. Considering the average percentage value of replies for C users in Figure 6.4a, the population percentage is 35% for C users and 75% for NC users.

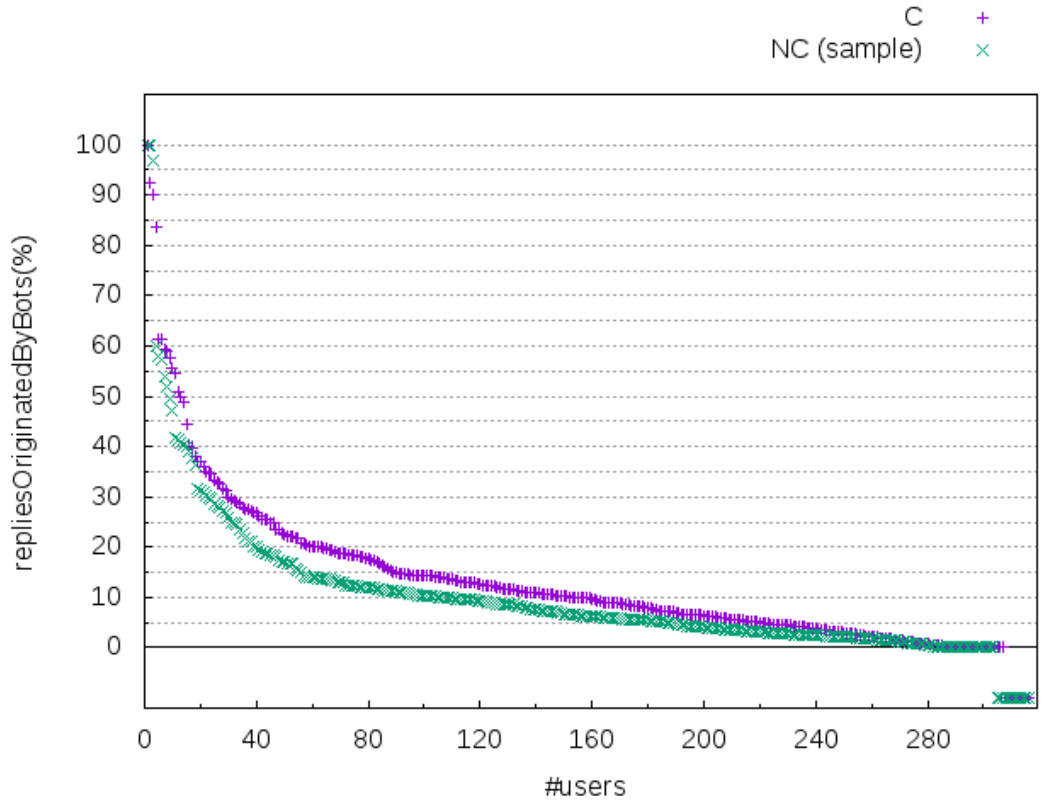
Like in the previous subsection, Figures 6.5a and 6.5b report the bar graphs related to the analysis of replies. The outcomes are very similar to those of the retweets analysis. Due to the low number of users, up to the group *[50,40[*, there is no a clear distinction between C and NC users (bot in Figure 6.5a and 6.5b). Instead, from *[40, 30[* to *[20, 10[*, the number of C users is increasing more and more compared to NC users. This holds at least until the *[20, 10[* group where NC users overcome C users.

Findings Similarly to what unveiled in the previous subsection, the replies analysis confirms that, on average, C users feature a higher percentage of replies to bots. Looking in more detail to the number of replies (the 'group analysis' in Figure 6.5, the groups with higher 'replies-to Bots' are not enough populated to allow relevant considerations. Instead, from the group named *[40, 30[* up to the *[20, 10[* one, we can notice a certain superiority of C-users in replying to tweets created by bots compared to NC users.

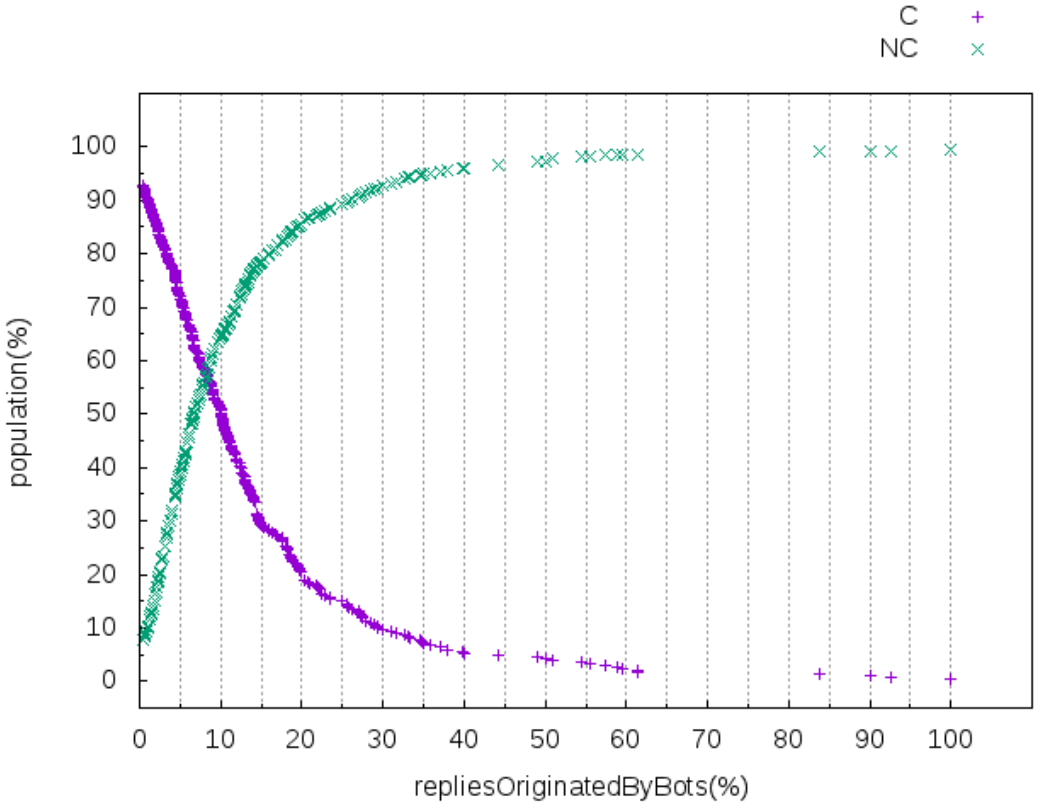
6.2.3 Quoted tweets

Figures 6.6 and 6.7 are concerned with results about quoted tweets.

Subfigure 6.6a, concerned with the percentage comparison of quoted tweets 'byBots' between C and NC users, shows that not always the purple points (C users) are above the green dots (NC users). This is different from the cases of replies and retweets discussed in the two previous subsections and might be due to the high presence of users that did not quote tweets at all, the outliers, in both populations, i.e., 125 C users (purple points under the y-axis zero) and 78 NC users (green points). From a numerical point of view, the average (in percentage) of quotes originated by bots is 18.8 (with a σ=24.96) for C users and 13.66 (with a σ=18.01) for NC users. Of course, the

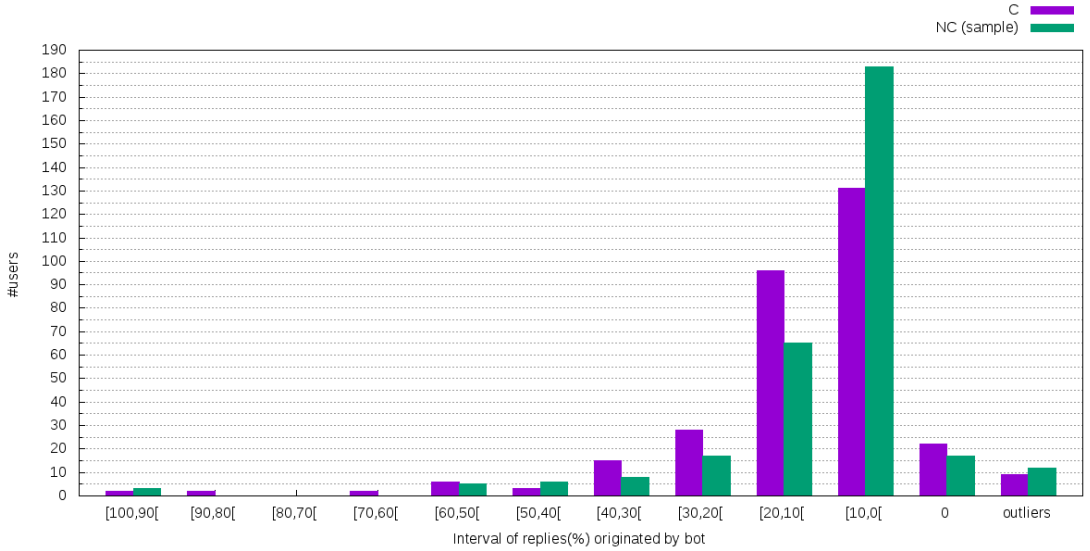


(A) Percentage of replies to bot's tweets posted by C and NC (sample) users.

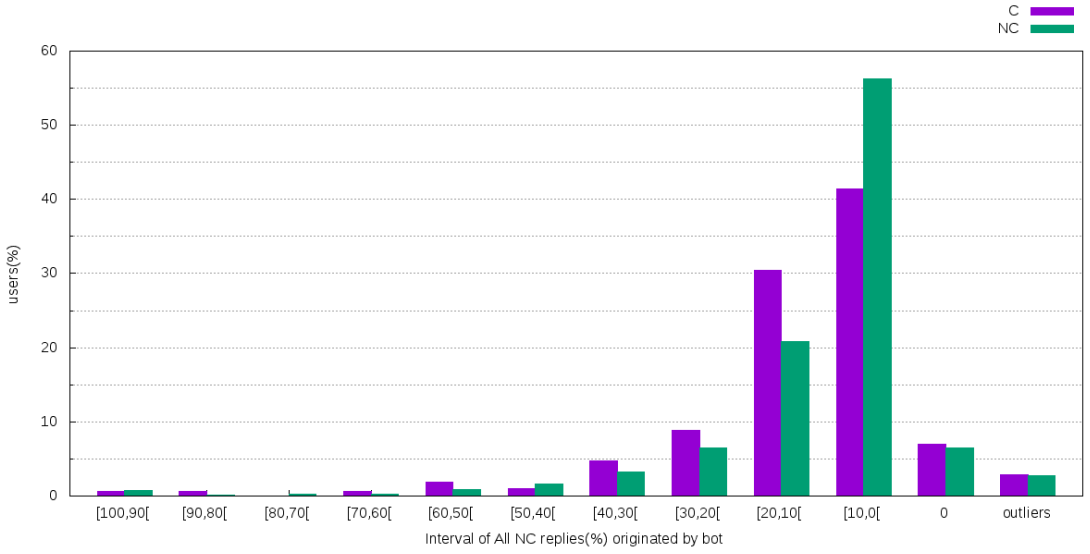


(B) % of populations w.r.t. the % of replies to bot's tweets

FIGURE 6.4: Comparative analysis between C and NC users w.r.t. the replies to bots' tweets.



(A) Deciles of Figure 6.4a

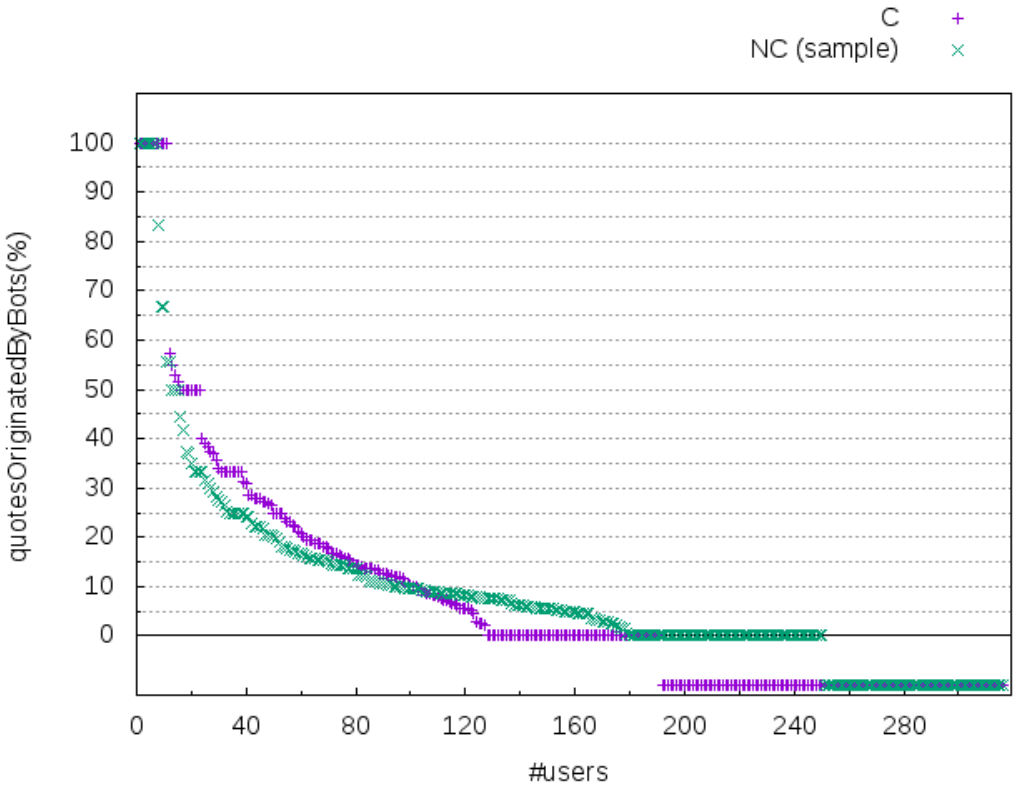


(B) Deciles of C and all NC users

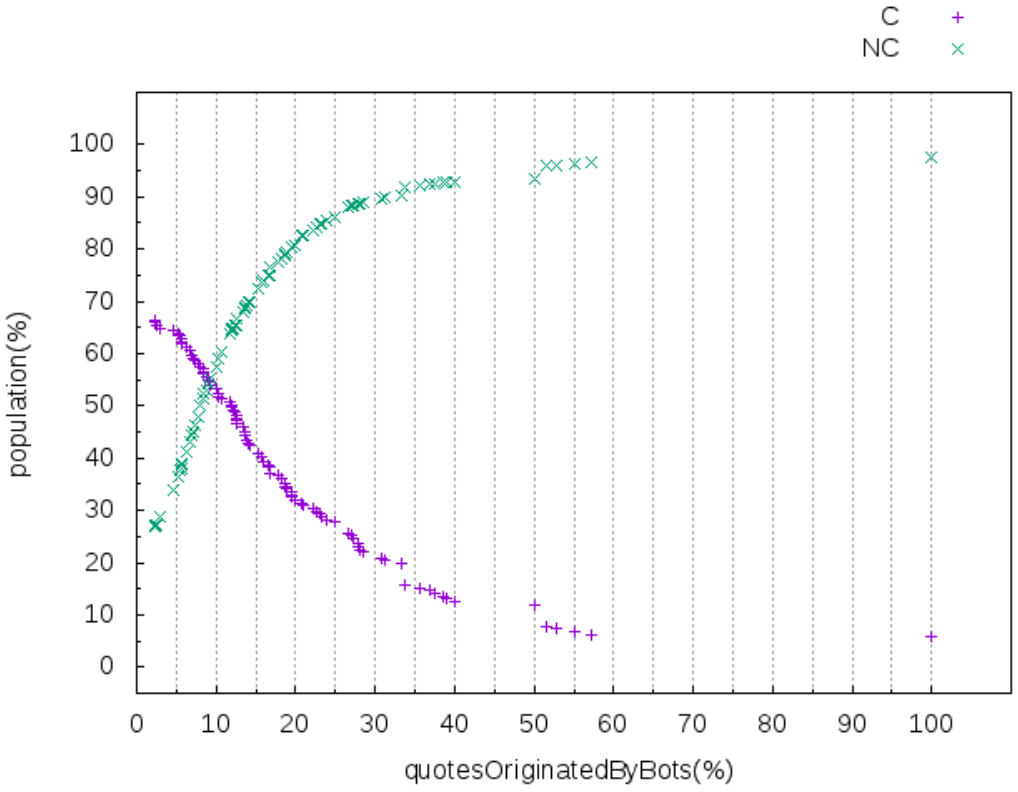
FIGURE 6.5: Analysis using deciles – C *vs.* NC users w.r.t. the replies to bots' tweets.

percentage of quotes originated by bots have been calculated over the total amount of quoted tweets, now considered no longer as retweets, as in Figure 6.1b, but as a different type of tweet. It is worth to remark that the users lying on the y-axis zero are those users whose quoted tweets have been originated by human-operated accounts only.

When considering Figure 6.6b, we found that the 44.5% of C users has a percentage of quotes 'byBots' $\geq$ 13.64%, while the 68.78% of NC users has a lower percentage. This represents the percentage of maximum coverage for both populations. If we would plot the same graph of Figure 6.6a, while considering only the users belonging to that percentages, we would obtain a clearer separation between C and NC users.

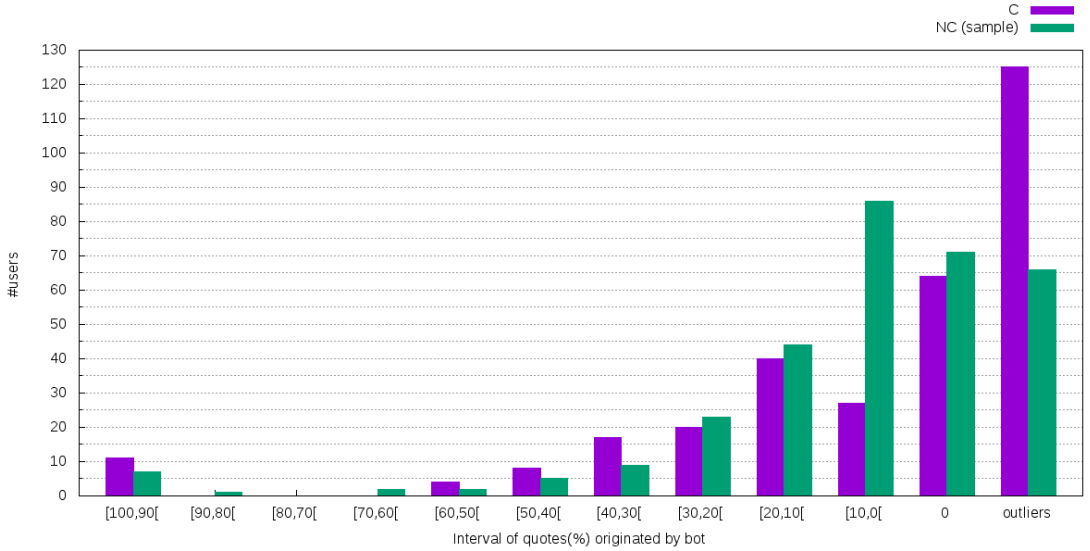


(A) Percentage of 'byBots'-quotes posted by C and NC (sample) users.

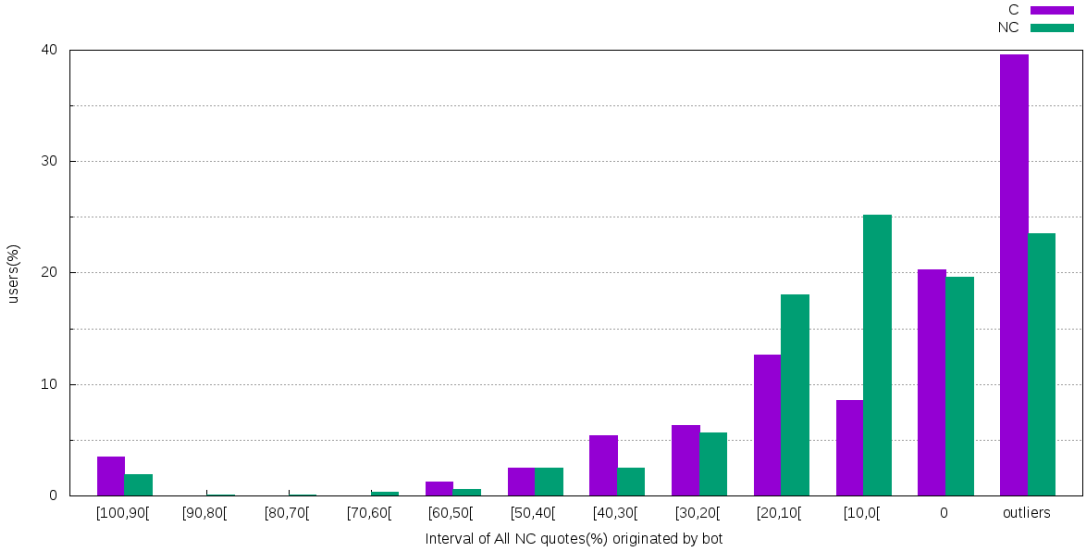


(B) % of populations w.r.t. the % of 'byBots'-quotes.

FIGURE 6.6: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes.



(A) Deciles of Figure 6.6a



(B) Deciles of C and all NC users.

FIGURE 6.7: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes

The behavioural analysis, related to quoted tweets, concludes with the bars in Figures 6.7a and 6.7b. Similarly to retweets and replies, also here two aggregation perspectives are reported, by grouping the C and NC users according to the number of quotes originated by bots. In Figure 6.7a, the sample of NC users is the same one we used for Figure 6.6a. While, in Figure 6.7b we consider all the NC users. When considering the whole population of NC users, we can notice that the comparison trend is preserved, and thus indicates the representativeness of the subset of 316 considered NC users (in Figure 6.7a). Finally, in both the subfigures of Figure 6.7, the bars falling in the last group (headed *outliers*) count the amount of users who did not quote any tweets. The bars belonging to the zero groups reports the amount of users whose quoted tweets have been originated by human-operated accounts.

Findings Looking at Figure 6.6a, we see that there is not a big difference between C and NC users, when considering the quotes of messages whose original tweets have been produced by a bot account. Qualitatively, the curves representing the two populations are rather close. However, from a quantitative point of view, a difference emerges when looking at the statistical descriptors of the two populations: although the standard deviations are similar, we have that, on average, C users quote more tweets originated by bots than NC users. Although investigating about the harmfulness of the original tweets is out of scope in this work, we cannot ignore the attitude of C users to give visibility to content that, being generated by bots, can be potentially malicious.

If we look more in depth at the behavioural activities of certain portions of the two populations (see Figure 6.6b), we can precisely identify the threshold value (percentage of bot-originated quoted tweets) that maximises the number of users to be considered for both populations.

Instead, in Figure 6.7 we can notice a dominance of C users in terms of amount of quotes whose tweets are originated by bots, in comparison with both a sample of NC users (Figure 6.7a) and all NC population (Figure 6.7b). The numerical dominance of C users is preserved until the group/decile *[40,30[*, which includes those users with a percentage of quoted tweets originated by bots $\geq$ of 40% and $<$ 30% (in Figure 6.7a); the same for the decile *[30,20[* in Figure 6.7b. The different trends in the two figures is mainly due to the high numbers/percentage of outliers. In fact, this represents the main limitation of the analysis performed on this particular kind of tweets.

6.2.4 Significance of the behavioral differences between C and NC users

As shown above (see Figure 6.2a for retweets, Figure 6.4a for replies and Figure 6.6a for quotes), there are behavioural differences between the C and NC populations. In this subsection, we aim at assessing whether these differences can be considered statistically significant. For this purpose, we perform statistical tests over groups of C and NC users, considering the same set of users.

Type of tweets	Kolmogorov-Smirnov Test of Normality	
	C (Res.)	NC (Res.)
Replies	×	×
Retweets	×	×
Quotes	×	×

TABLE 6.2: Kolmogorov-Smirnov test (Test of Normality).

Type of tweets	T-Test (α=0.05)			ANOVA (α=0.05)		
	Res.	t-value	p-value	Res.	f-ratio	p-value
Replies	✓	3.001	0.001	✓	9.04942	0.002738
Retweets	✓	3.190	0.001	✓	10.17804	0.001496
Quotes	✓	2.472	0.138	✓	6.11248	0.13812

TABLE 6.3: Parametric Statistical tests: T-test and one-way ANOVA.

Table 6.2 shows the results of the first statistical tests known as the Kolmogorov-Smirnov's test [101] (also know as *Test of Normality*). It is a non parametric test that, given a certain number of observations (in our case the percentages 'byBots'), checks whether such observations are normally distributed. Hence, the *null*-hypothesis claims that "there is no significance in data to state they are following a normal distribution". We perform this first test both on C and NC (sample) users. If the test is successful, then we can rely on the outcomes obtained by performing parametric statistical tests on C and NC users' data; in particular the T-test [158] and one-way Analysis of Variance [81] (ANOVA). Unfortunately, as indicated in the column headed *Res.*, both populations did not pass the test (symbol ×). This means that there is not enough grounds to reject the (aforementioned) null hypothesis. Therefore, information obtained by parametric statistical tests is thus useless in our situation. Anyway, just for sake of curiosity and completeness, we also reported in Table 6.3 the outcomes obtained by conducting both parametric tests for each type of tweets. However, these outcomes will not be considered further.

Type of tweets	Mann-Whitney (α=0.05)			Kruskal–Wallis (α=0.05)		
	Res.	z-score	p-value	Res.	H-value	p-value
Replies	✓	3.37056	0.00038	✓	11.36	0.00075
Retweets	✓	3.3	0.00048	✓	10.89	0.00097
Quotes	×	-1.20349	0.11507	×	1.5	0.22017

TABLE 6.4: ANOVA and Mann-Whitney (not parametric) tests.

Because of the outcomes of the Kolmogorov-Smirnov test (see Table 6.2), we prefer to rely on non-parametric statistical tests. In Table 6.4, we report the outcomes of two well-known non-parametric statistical tests which correspond to the non-parametric version of T-test and ANOVA; precisely, the Mann-Whitney [109] and Kruskal–Wallis tests [91]. For both of them, the test is reputed to be successfully passed if there is enough grounds to reject the *null* hypothesis. Roughly, in both tests, the *null* hypothesis states that "there is no difference in means" (of 'byBot' content) between the different populations (in our case C and NC users).

As we can see in Table 6.4, only two types of tweets (i.e., replies and retweets) successfully pass both tests; instead, for quoted tweets, it is not possible to reject the null hypothesis.

These results suggest that when replies and retweets are considered, C users interact more with bots than NC users and this behavioural difference is not due to chance.

6.2.5 Retweets and quoted tweets: an aggregated view

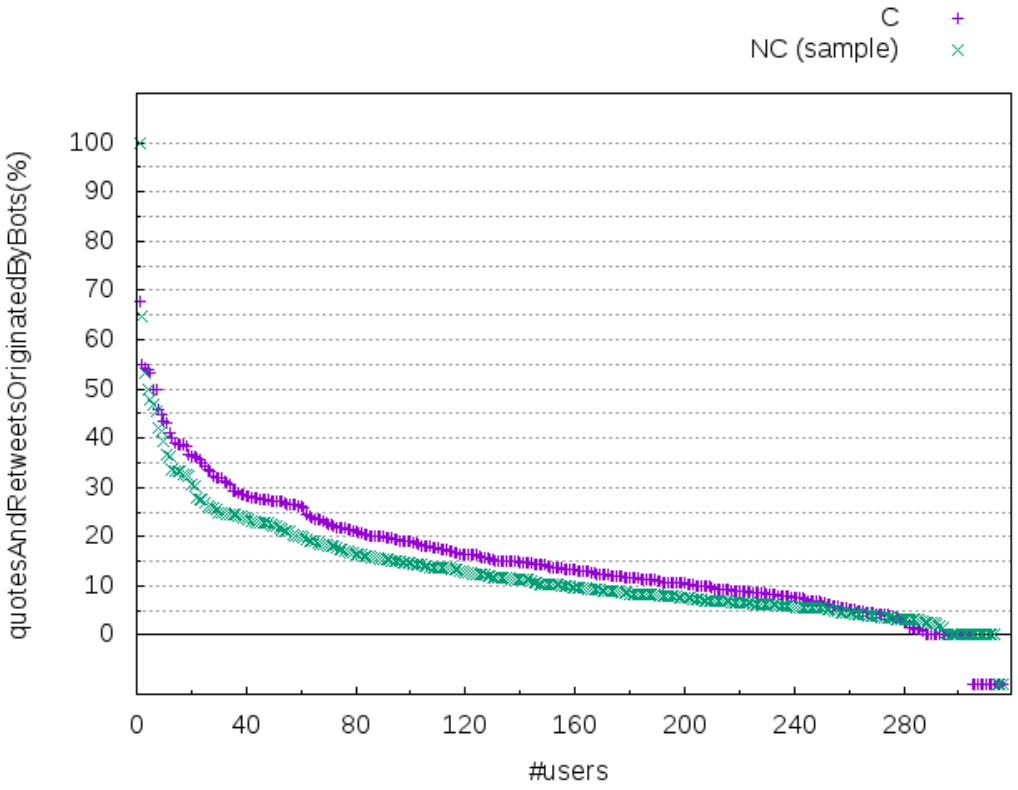
Now, we further investigate the behavioural differences between C and NC users when quoted tweets are considered as retweets, exactly as done in the coarse-grained analysis at the beginning of this section (see Figure 6.1). Like in Sections 6.2.1-6.2.3, we report the outcomes concerned with the bounced 'byBots' tweets.

In Subfigure 6.8a, we compare the percentage of content 'byBots' that C and NC users quote or retweet. Almost all points related to C users (purple dots) overpower the NC users' ones (green dots). Unlike the quoted tweets analysis in Section 6.2.3, here the number of outliers is more similar to the retweets case (Section 6.2.1); precisely, we have 12 C users and 6 NC users as outliers. The average of content 'byBots', related to C users, is 16.22% (σ=11.6%), while the one corresponding to NC users is lower, 13.41% (σ=10.48).

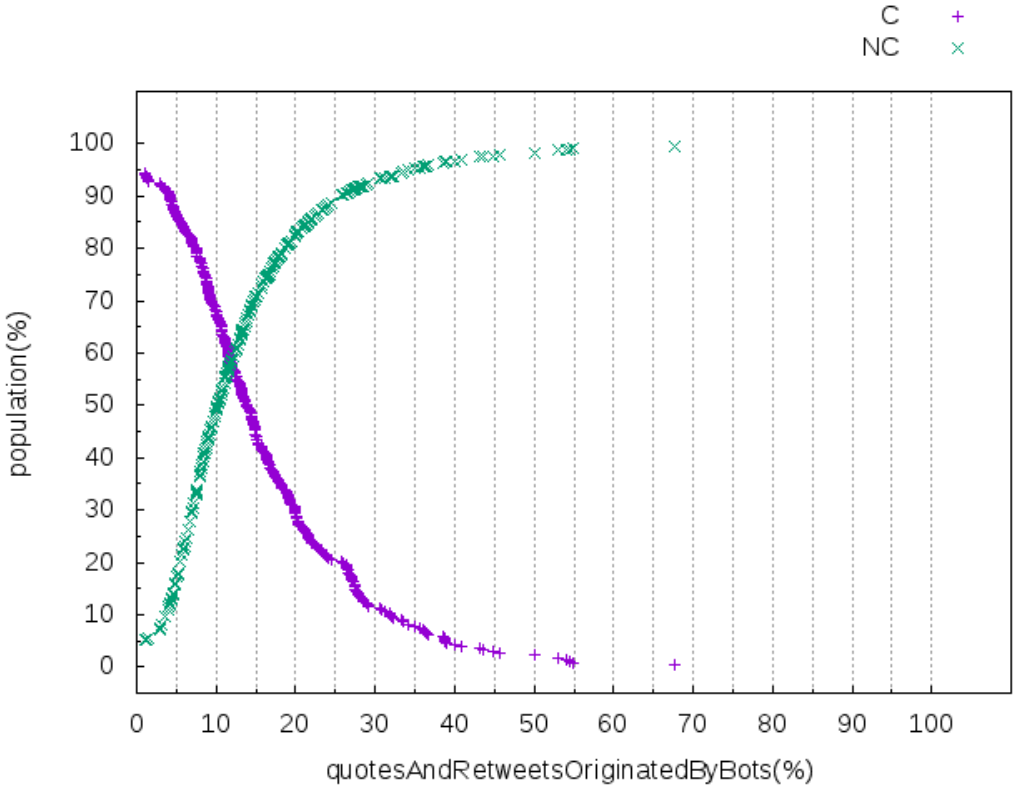
Considering Subfigure 6.8b, we found that the maximum percentage of population coverage is associated with 19.01% of content 'byBots'. In particular, we can see that the percentage of C users, whose 'byBots' content exceeds such threshold, is 32%; while, 81% of NC users have in their timeline less than 19% of the tweets originally posted by bots.

The subfigures in Figure 6.9 report the bar graphs related to the aggregation perspective (deciles). Figure 6.9a and 6.9b are very similar, and this still confirms that the NC sample is representative of our NC population. We can immediately see that up to the *[10,0[* group, both the number (Fig. 6.9a) and the percentage (Fig. 6.9b) of C users exceed that of NC users. In this case, due to the low number of outliers, we have more observations to analyse reinforcing the reliability of the claims derived from these last charts with respect to the case where quotes have been considered in isolation (see Seciont 6.2.3).

We will do the statistical tests of Section 6.2.4 also for this new behavioural analysis. The normality tests (Kolmogorov-Smirnov) fail again for both C and NC populations, demonstrating the non-uniform data distribution and thus the lack of information that a parametric test can give for statistical purposes. Instead, both T-test and ANOVA successfully passed (T-test: t-value= 3.145 and p-value= 0.001 – ANOVA: *f-ratio= 9.894* and *p-value = 0.002*). For quoted tweets considered as retweets, we repeat the Mann-Whitney and Kruskal-Wallis tests, both of them performed with a significance level of

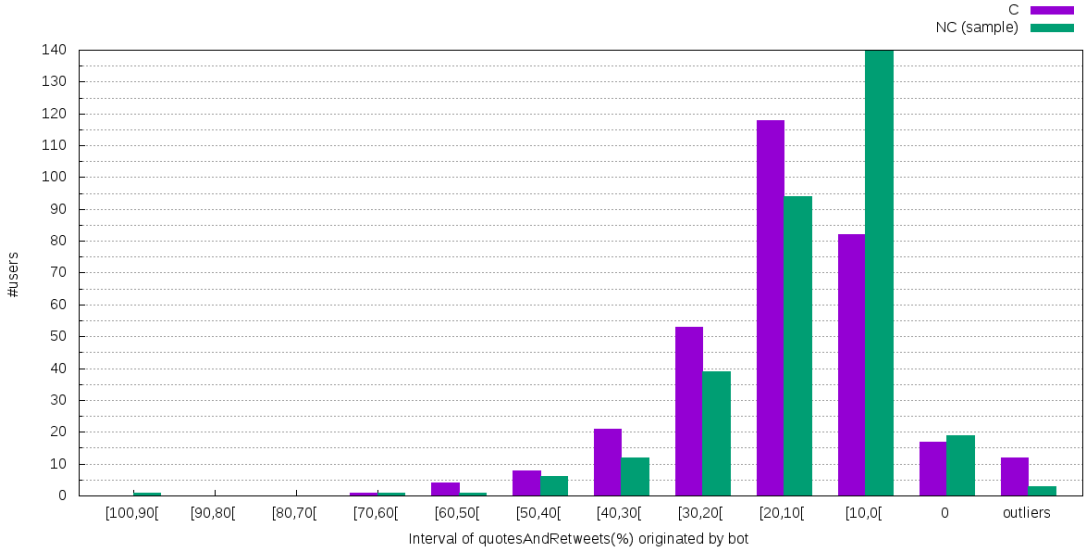


(A) Percentage of 'byBots'-quotes and retweets (jointly) posted by C and NC (sample) users.

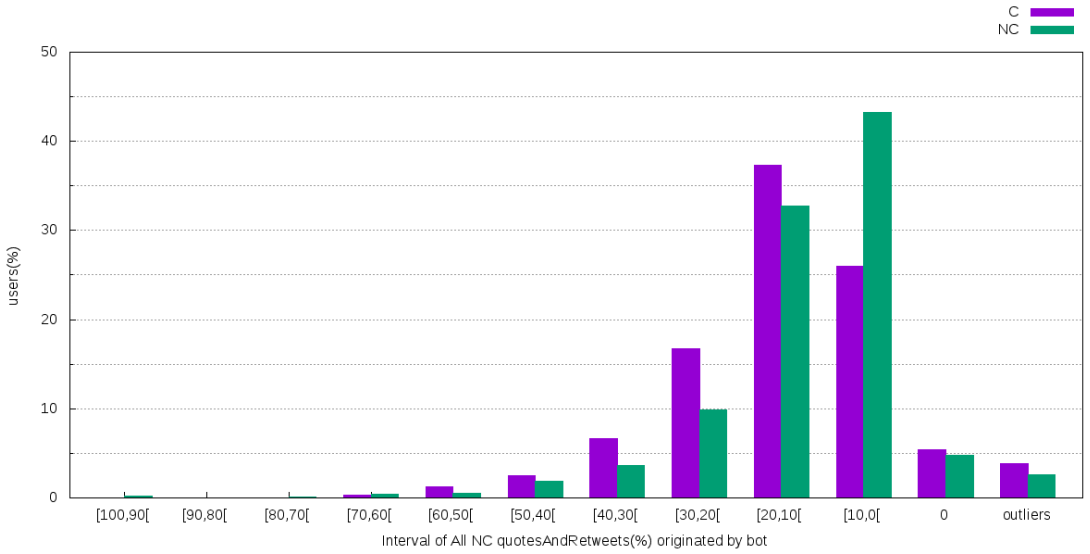


(B) % of populations w.r.t. the % of 'byBots'-quotes and retweets (jointly).

FIGURE 6.8: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes and retweets (jointly).



(A) Deciles of Figure 6.8a.



(B) Deciles of C and all NC users.

FIGURE 6.9: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes and retweets (jointly).

$\alpha = 0.05$. In the former, we get a z-score $= 3.498$ and a p-value $= 0.0002$; in the latter, we obtain an H-value $= 12.239$ and a p-value $= 0.0005$.
In this case, both the non parametric tests reject the null hypothesis, unlike the case of considering quoted tweets by themselves.

Findings Unlike the case where quoted tweets were observed in isolation, here, thanks to the support of statistical tests, we can say that there is a difference between the populations of C and NC users, and that such a difference is statistically significant.

6.3 Further analysis

In the following, we extend the analysis to two additional sets of C users, namely: *cut946* and *cut1030*, as they have been called in Section 4.3.2. Here, we summarise the results in a tabular mode. The complete graphs are in Appendix B and referred in the tables. We assign priority to the analysis related to the percentage of content 'byBots' (subfigures *a)* in previous subsections) and population coverage (subfigures *b)* in previous subsections). The graphs related to the groups analysis (deciles) can be found in Appendix B, for the sake of completeness.

We start to show the results of this analysis by considering 443 human-operated accounts as C users (namely, the set called *cut946* in Section 4.3.2). The number of non-credulous users is 2,395.

Types of		C			NC* (sample)			Deciles (Fig.)	
tweets	Fig.	μ (%)	σ (%)	out.(#)	μ (%)	σ (%)	out.(#)	NC*	NC
Retweets	B.1a	16.44	12.41	13	12.37	10.24	12	B.2a	B.2b
Replies	B.3a	12.98	14.33	11	10.08	11.85	13	B.4a	B.4b
Quotes	B.5a	18.15	24.27	162	11.88	16.51	107	B.6a	B.6b
QuPlusRw	B.7a	16.16	12.34	13	12.16	10.08	11	B.8a	B.8b

TABLE 6.5: Mean, standard deviation, and # outliers per content originated by bots – 443 C users *vs.* 2395 NC users (*cut946*).

Table 6.5 reports, for both C and NC users, the average, the standard deviation and the number of outliers related to content 'byBots' posted by both C and NC users. It is worth considering that, like the analyses in previous sections, the NC users are a (representative) sample of the whole population. Similarly to the previous case, when the set of C users consisted of 316 human-operated accounts, here we can appreciate that, regardless of the type of tweet, the values corresponding to the averages of 'byBots' content of C users (column μ in Table 6.5) overcome those related to (the sample of) NC users. This confirms a greater attitude of C users to spread content originated by bots, with respect to NC users.

Types of tweets	_'byBots'(%) max	C (pop %) $\geq$ max	NC (pop %) $<$ max	Fig.
Retweets	15.59	42.56	71.62	B.1b
Replies	13.81	32.21	76.68	B.3b
Quotes	3.51	65.84	29.82	B.5b
QuPlusRw	18.82	31.18	79.96	B.7b

TABLE 6.6: Populations coverage analysis (resume) – *cut946*

Like for the population coverage analysis performed in previous subsections, in Table 6.6 we perform the same analysis on *cut946*. For this analysis all the 2,395 NC users are

taken into account. In general, by looking at the percentages referred to the C and NC users, we can appreciate a good coverage of C users w.r.t. the NC users population.

We conclude our further analysis by experimenting with the largest set of C users (namely, *cut1030* in Section 4.3.2) that includes 502 accounts and 2,336 NC users.

Types of		C			NC* (sample)			Deciles (Fig.)	
tweets	Fig.	μ (%)	σ (%)	out.(#)	μ (%)	σ (%)	out.(#)	NC*	NC
Retweets	B.9a	16.15	12.10	13	13.85	12.26	21	B.10a	B.10b
Replies	B.11a	12.81	13.86	12	11.25	14.30	18	B.12a	B.12b
Quotes	B.13a	17.23	23.16	174	14.49	19.59	135	B.14a	B.14b
QuPlusRw	B.15a	15.90	12.02	13	13.56	12.13	19	B.16a	B.16b

TABLE 6.7: Mean, standard deviation, and # outliers per content originated by bots – 502 C users *vs.* 2336 NC users (*cut1030*).

Table 6.7 shows that C users post (on average) more content originated by bots than NC users. We do notice a decreasing gap (in terms of difference between averages) between the μ(C) and μ(NC) when compared to the *cut946* case. Moreover, by better looking at (and comparing) the μ(C) values of the *cut946* case with the current μ(C) values, we can see that the values decrease, but this decrease is not so evident, because of the few users (69) who have become part of the C users.

Types of tweets	‘byBots’(%) max	C (pop %) $\geq$ max	NC (pop %) $<$ max	Fig.
Retweets	15.59	41.92	71.85	B.9b
Replies	6.81	65.11	52.12	B.11b
Quotes	10.77	48.48	60.97	B.13b
QuPlusRw	11.84	55.83	59.24	B.15b

TABLE 6.8: Populations coverage analysis (resume) – *cut1030*

As in the previous case (Table 6.6), by looking at Table 6.8, we can notice (3rd column) that a considerable part of C users bounce more ‘byBots’ content than a high percentage of NC users (last column) for each type of tweet. This

Findings Regardless of the number of C users, they resulted to bounce more ‘byBot’ content than NC users. However, when enlarging the number of C users, the gap (in terms of mean of byBot content) with respect to NC users decreases; such a trend further strengthens the validity of our method of singling out credulous users. In fact, if in the last two cases the gap between C and NC users increased rather than narrowed, it would have meant that the set of 316 users, identified of being credulous and experimented in Section 6.2, could not be considered well-defined. This may be due to the presence of some users, considered as being not credulous, but to be so in *cut9476* and *cut1030*, that exhibit behaviours that (by definition) were attributed to credulous ones (i.e., bouncing

content created by bot). Consequently, even the method used to single out the credulous users (in Chapters 3 and 4) was to be considered fallacious.

6.4 Discussion

This chapter investigated the harmfulness of C users on social media for spreading potentially malicious content from bots. To identify behavioural differences, we compared C and NC users behaviour looking at different types of tweets, i.e.,: *pure tweets* or self-originated tweets, replies and retweets. Except for the *pure* tweets case, where C users proved to be much more prolific than NC users on average, the analysis (defined as coarse-grained) did not show significant differences between these two populations of human users. To this purpose, a fine-grained analysis has been conducted. Precisely, we inspected the nature (human or bot) of the originators of contents. In this analysis retweets and quoted tweets have been examined separately. A more marked difference was observed, in fact on average C users resulted to bounce more bot originated content than NC users.

To ensure that this difference is statistically significant, non-parametric tests have been performed to assess whether these behaviours of C and NC users are indeed different. These tests were conducted for each type of tweets and confirmed the statistical significance, except for quoted tweets case. Therefore, the analyses have been repeated, by considering retweets and quotes together, and statistical tests proved that C and NC populations behave differently.

Similar to previous chapters, here we also extended our investigation to larger sets of credulous users, and on average C users resulted to bounce more content from bots than NC users, but with smaller differences (in terms of average between C and NC users).

This investigation gives us sufficient ground to provide an answer to the forth research question, reported below for the convenience of the reader.

> **RQ4** – Is it enough to compare the different types of activities (i.e., retweets, quoted tweets, replies and posting new content) between credulous and not credulous users to significantly differentiate them? Can bot-followees influence, in terms of content production, the activities of credulous users more than not credulous ones? How to measure the effectiveness of such an influence? Do credulous users bounce bots' content? And to what extent with respect to not credulous users? (see Chapter 6)

ANSWER *–From the high similarity of the statistical descriptors (see Fig. 6.1), related to the credulous and not credulous users and calculated on the different types of tweets, we can deduce that it is not possible to single out significant dissimilarities between these two categories of users via a coarse-grained analysis (except for the pure tweet case). The influence of bots-followees on credulous users, in terms of manipulation of the disseminated content, emerges from a more in-depth analysis conducted on the authors of the tweets (subsequently bounced by other human-operated accounts, both credulous and not-credulous users). Thanks to this fine-grained analysis, we can appreciate behavioural differences between the two populations highlighting that, on average, credulous user bounce more bot-originated content than not credulous ones.*

Chapter 7

Credulous Users and Fake News

7.1 Introduction

Inspired by the findings from the previous chapter, where we provided the evidence about credulous users' involvement in spreading bot-originated contents (therefore potentially malicious), in this chapter the focus is on the relationship between *credulous* users and fake news. Investigating such a relationship can provide us, not only with further evidence about the (harmful) contribution of credulous users to the dissemination of malicious content, but it can be seen as an indirect and alternative way (credulous users centered) to deal with mis-/dis- information.
Usually the problem of fake news is addressed in a 'direct' way, i.e., by trying to establish whether a news is true or not. On this direction, several approaches have been developed to counteract the spread of this phenomenon [20, 148]; for instance, by using Natural Language Processing (NLP) techniques [126] to analyse the actual content in messages. Unfortunately, despite the great progress made by the scientific community on this matter, the current fake news detectors lack in the recognition effectiveness [151]. Given that, instead to apply a fake news detector to the content published by the Twitter accounts in our dataset (i.e., *Humans2Consider*), we prefer to consider as starting point a (publicly available[1]) dataset of news already annotated as fake or real [151]. The authors of such news will be analysed (to distinguish the credulous from not credulous users) and the proportions of fake news spread by these two categories evaluated (e.g., the number of tweets containing fake or real news).

With this chapter, we takes in charge the issues related to the fifth research question. It is worth to notice that the obtained findings have been published in [10].

[1] https://tinyurl.com/uwadu5m

7.2 Experimental setup

In this section, we start by describing the employed data (their source and meaning), then the adopted approach is presented. Finally, we explain the target of our analysis and the considered perspectives to derive our findings.

7.2.1 Dataset

As mentioned before, our investigation starts by considering a publicly available dataset of fake news, called *FakeNewsNet*[2] [151–153]. For each item the following information are provided: a unique identifier (*id*), the publisher (in *url* form), the content of the news (*title*), a list tweets (as Twitter *ids*) containing the news and the information about its "veracity" (fake or real). To label the news, the authors in [151] used two fact-checking websites: *PolitiFact*[3] and *GossipCop*[4]. In the former, fact-checking was performed by politics experts (e.g., journalists) labelling news as fake or real. In the latter, a numerical scores was assigned to news to indicate their reliability, ranging from 0 (fake) to 10 (real).

		News		**Tweets**	
		Original	Retrieved	Original	Retrieved
Politic	Fake	432	392	165,356	141,421
	Real	622	407	417,072	357,655
Gossip	Fake	5,323	5,135	598,299	518,502
	Real	16,817	15,759	881,627	812,719

TABLE 7.1: FakeNewsNet Dataset: original and retrieved content

Table 7.1 outlines the dataset's details. The original dataset contained 432 fake and 622 real political news (see in row *Politic* the column *Original*). However, on Twitter we were only able to find tweets of 392 (91%) fake and 407 (65%) real news (column headed *Retrieved*); while 9% of the false news and 35% of the real news was no longer available. The number of *tweets* (column titled *Tweets*) containing such news was initially of 165,356 on fake and 417,072 on real news; but we could find only 141,421 (86%) tweets related to fake news and 357,655 (86%) tweets related to real ones. The untraceable

[2]FakeNewsNet Dataset: https://tinyurl.com/uwadu5m
[3]https://www.politifact.com/
[4]https://www.gossipcop.com/

tweets are 14% concerning both real and fake news. The numerical mismatch between the *original* and the *retrieved* data is almost certainly due to deletion.
Regarding the other topic, (row headed *Gossip*), out of 5,323 fake and 16,817 real news, we got 5,135 (96%) and 15,759 (94%) news, respectively. Not retrieved news are 4% and 6% for true and false news, respectively. Such news are contained in 1,331,221 tweets, 518,502 (87% but *n.a.* 13%) related to fake news and 812,719 (92% but 8% n.a.) containing real news. Obviously, in our study, we only use retrieved data.

7.2.2 Approach

To shed light on the credulous users' involvement in fake news dissemination, we single out three sequential tasks: (i) tweets' authors identification, (ii) distinction between automated (bots) and human-operated authors, and (iii) distinction between *credulous* and *not-credulous* users among the human-operated authors.

Tweets' authors identification

We aim to identify the Twitter accounts that published the tweets listed in *FakeNewsNet* dataset (referred as authors). Starting from the tweets IDs and using Twitter API[5], we collected 1,731,422 tweets out the *original* list of almost 2 million. It might be worth noting that some tweets contain more than one news, and thus some tweets are counted more than once in Table 7.1. This explains the numerical mismatch between the collected tweets and retrieved ones (sum of the values in 4th column).
At the end of this phase, in addition to tweets' data, we collected the profile's data[6] of 536,513 Twitter accounts, i.e., the authors of all the retrieved tweets.

Bot detection

Here the goal is to distinguish, among the set of authors spotted out in the previous phase, between human-operated accounts and bots. To this purpose, we used a bot detector (i.e., a decision model able to recognise automated accounts) introduced in our previous work [12] and described in Section 4.2.2. We recall that the classification model is based on Random Forest [23], achieving an *accuracy* (instances correctly classified) of 98.41% and an area under the ROC curve (AUC) of 1.00. It relies on a set of 30 features (called *ALL_features*) obtained by combining the feature sets of *Botometer+* and *ClassA-*, see Chapter 4 for further details. For each user, we retrieve *Botometer+* features

[5]Twitter API libraries: https://tinyurl.com/rfte3k2
[6]Twitter User Object: https://tinyurl.com/y5s5kpuw

considering the *timeline* (the list of published tweets) and its *mentions* (the tweets that mention the user). *ClassA*-'s, instead, relies on users' profile data to determine their features[6]. This way, 479,569 authors have been classified as human-operated accounts.

Credulous classification

This task aims at singling out *credulous* users among the human-operated authors. To this purpose, a refined version of the approach presented in [12] has been adopted.

Instead of classifying authors using the best classifier (trained with just one fold in [12] – see Section 4.2.4), we use all the eight *credulous* classifiers, and classification performances are reported in Table 7.2.

		evaluation metrics					
Fold	*alg*	*accuracy*	*prec.*	*rec.*	*F1*	*MCC*	*AUC*
1	OneR	98.26	0.98	0.98	0.98	0.97	0.98
2	OneR	95.73	0.96	0.96	0.96	0.92	0.96
3	OneR	94.15	0.95	0.94	0.94	0.89	0.94
4	JRip	90.67	0.92	0.91	0.91	0.83	0.89
5	RepTree	91.93	0.93	0.92	0.92	0.85	0.90
6	OneR	90.35	0.92	0.90	0.90	0.82	0.90
7	OneR	90.93	0.92	0.91	0.91	0.83	0.91
8	OneR	96.65	0.97	0.97	0.97	0.93	0.97

TABLE 7.2: The eight Credulous Classifiers

This way, each author is classified by means of the classifier trained on the *fold* most "similar" to it. Hence, for each human author singled out in Section 7.2.2, the distance between author's feature representation and the centroids of each *fold* is computed. The author is then classified using the classifier trained on the fold whose centroid is closest to it. Classifiers selection is performed with a specific tool we have implemented; and 350,622 human authors have been classified as *credulous* users.

7.2.3 Investigation targets

At this stage, we have all the information to investigate on a potential relationship between fake news and *credulous* users. Three different perspectives are considered. First, we look for numerical differences between the amount of fake and real news produced/diffused by the three categories of user/author (namely, *credulous*/*not-credulous* and bots) and on both news' topic (i.e., gossip and politics). Second, we are interested to compare the quantity of fake/real tweets, i.e., the tweets containing a fake/real news,

by looking at first bots and humans and then *credulous* and not-credulous[7]. Third, we quantify the *authors' level of involvement* in fake news spreading/production by counting, for each category, how many of them are authors of tweets containing: at least one fake news, at least one real news, only fake news and only real news.

7.3 Experimental results

Table 7.3 shows the results obtained for bot and *credulous* detection (Section 7.2.2). First column lists the different types of Twitter users investigation; in the first macro-row the difference is based on the "automation" of an account (human or bot). In the second macro-row it is reported the numbers of human-operated accounts labeled as *credulous* or *not-credulous* users by using the classification approach described in Section 7.2.2. Each column reports the number of users which tweeted about a certain topic with the exception of the last one (namely *Union*) where the total number of users for each category of account (e.g., bot and human) is reported. It is worth specifying that, for each user category (corresponding to the single row in the table), the sum of the values between the first and second column (i.e., *Politic* and *Gossip*) does not coincide with the value reported in the third one since some accounts have tweeted the news in both topics. Note that our classifier was not able to classify 396 accounts due to lack of information (empty answer) from the *Botometer* web service.

	Politic	**Gossip**	**Union**
#Bot	27,137	34,160	56,548
#Human	256,561	247,113	479,569
#Credulous	185,196	178,715	350,622
#Not-Credulous	71,365	68,398	128,947

TABLE 7.3: Detectors outcomes

Table 7.4 shows the results obtained by a *news*-centered perspective. Precisely, here we observe the amount of news that each category of users, i.e., credulous/not-credulous users and bots (1st column) covers with their tweets over the total number of retrieved news. By looking the data under this perspective, we aim to assess a sort of news' interest, grouping them by the topic (*Politic* or *Gossip*). It is very important to keep in mind that this perspective does not take into account how many times a specific news

[7]To avoid that the numerical unbalancing between fake and real tweets (e.g., see in Table 7.1 political *retrieved* tweets) may lead to inaccurate observations, we will also consider a randomly selected subset, among the set of real-news tweets, with the same number of fake-news tweets.

	Politic		**Gossip**	
	Fake	*Real*	*Fake*	*Real*
Credulous	373	364	4,121	14,486
NotCredulous	361	366	4,768	13,418
Bot	350	332	4,470	15,050

TABLE 7.4: Users' topic coverage by their tweets

has been bounced (e.g., how many tweets it received) but just if it has been tweeted. For instance, regardless to the topic or the veracity, let consider two different news (*newsA* and *newsB*), if *newsA* has been tweeted 100 times and *newsB* just 2, both contribute to count just one in the table. The 2nd and the 3rd column, (macro-column headed *Politic*), indicate the amount of political news, respectively fake and real, that users have used in their tweets. The 4th and 5th column, instead, indicate the number of real and fake news about the gossip, respectively.

With their tweets, credulous users cover 373 fake and 364 real political news. The number of news covered by *not-credulous* users is 361 fake and 366 real political news. For the sake of completeness, the same information is reported for bots too; the amount of news covered by them is 350 fake and 332 real political facts. Taking into account the number of *retrieved* political news (see Table 7.1), and by comparing credulous *vs.* not credulous users, we can see that the tweets produced by *credulous* users cover 95% of the *retrieved* fake news concerned with politics, while those produced by *not-credulous* cover 92%. Instead, for non fake news, *credulous* users "talk" about the 91% of the total *retrieved*, which is almost the same of *not-credulous* users.

When considering *Gossip* news, starting from a retrieved set of 5,135 fake news, we have the following situation: 4,121 published by *credulous* users, 4,768 by *not-credulous* users and 4,470 by automated accounts (bots). For real news, we retrieved 15,759 tweets of which 14,486 come from *credulous* users, 13,418 from *not-credulous* and 15,050 from bots. We can see a decreasing about news percentage coverage of credulous users in fake gossip; indeed, the percentage of fake news covered by *credulous* users' tweets is of 80%, while the one related to *not-credulous* users' tweets is of 93%. This reduction can be explained either as a lesser interest that credulous users have in fake gossiping or as their greater caution with respect to fake policy news. On the other hand, *not-credulous* users "talk" of 86% of the *retrieved* true news, while the coverage of *credulous* users is of 92%. The fact that credulous users seem to diffuse a higher percentage of real news vs not credulous users may appear a counter-intuitive result. However, we have to stress that fake news represent the problem, not the real ones. Credulous users result not only interested in fake news but also in real ones.

Below, we show the results related to the *tweeting* perspective, from which we got the most relevant findings of our analysis. By this perspective, it is possible to investigate how many times fake news have been bounced and by which kind of users. Tables 7.5 and 7.6 provide a more detailed view of our experiments and shed light about the level of involvement of credulous users in misinformation activities. For each category of users (i.e., credulous, not credulous users and bots), it is reported: the number of tweets containing fake news (column *FN*) and real news (column *RN*) for both politics (Table 7.5) and gossip (Table 7.6). In both tables, the 4th column (called RN_{rnd}) and 5th column (called RN^*_{rnd}) are introduced to mitigate the unbalance between fake and real tweets, in accordance with the discussion in Section 7.2.3. With RN_{rnd} we denote a subset of *RN* whose entries have been randomly selected, from the *original* list of tweets of Table 7.1, in order to have $|RN_{rnd}| = |FN|$. While in RN^*_{rnd}, tweets are taken from the *retrieved* list of Table 7.1. The values in parenthesis express the percentage of fake news over the total fakes (according to the topic) with respect to each kind of users.

	FN (%)	RN	RN_{rnd}	RN^*_{rnd}
Tot.	165,356†	417,072	165,356	141,421
Bot	19,888 (12.03†)	45,924	18,120	18,013
n.a.	23,935 (14.47†)	59,417	23,519	0
Human	121,533‡ (73.50†)	311,731	123,717	123,408
Credulous	84,362 (69.41‡)	197,454	78,528 $\downarrow_{FN}$	77,994 $\downarrow_{FN}$
Not Credulous	37,171 (30.59‡)	114,277	45,193 $\uparrow_{FN}$	45,414 $\uparrow_{FN}$

Table 7.5: Number of tweets about political fact

	FN (%)	RN	RN_{rnd}	RN^*_{rnd}
Tot.	598,299†	881,627	598,299	518,502
Bot	116,398 (19.45†)	486,907	330,425	310,810
n.a.	79,797 (13.34†)	68,908	46,552	0
Human	402,104‡ (67.21†)	325,812	221,322	207,692
Credulous	244,690 (60.85‡)	209,579	142,246	133,518
Not Credulous	157,414 (39.15‡)	116,233	79,076	74,174

Table 7.6: Number of tweets about gossip fact

Table 7.5 presents the information related to the tweets on political fact. Almost 20k fake tweets have been produced by bots, more than 121k fake tweets have been produced by human-operated accounts; while for 24k fake tweets it has not been possible to retrieve their information from Twitter (row named *n.a.*). When considering *human* users, we can see that the number of fake tweets (*FN*) published by *credulous* users (i.e., 84k)

overcomes the number of those published by *not-credulous* ones (i.e., 37k).
Although the tweets in *RN* (total set) are more numerous than *FN*, the number of FN tweets authored by *credulous* users overcomes (69.41%) that of not-credulous (30.59%) ones.But, because the $\#RN$ humans' tweets are almost three times (∼311k) the $\#FN$ ones (∼121k), looking to the values related to RN_{rnd} and RN^*_{rnd} columns can led to a fairer comparison. In fact, we can note that the number of real tweets published by *credulous* users (in 4th and 5th column) are similar but in any case lower than in *FN* (see $\downarrow_{FN}$ in Table 7.5).
When bots are concerned, despite *RN*'s value seems higher then *FN*, the amount of fake and real tweets is more or less the same (see the 4th and 5th column). Regardless of news' veracity, from Table 7.5, the low number of tweets made by bots compared to human-operated accounts attracts attention; just 12.03% of FN and 10.99% of RN. In our opinion, these low values are mainly due to the disproportion between the amount of bots and humans (see Table 7.3), that confirms the statement in [166], i.e., the percentage of bots in Twitter is estimated to range from 9% to 15% (10.52% in our case).

Switching to the case of tweets containing gossip news (Table 7.6), we can notice that, like Table 7.5, also here there is a superiority in tweet's production by *credulous* users. In particular, by focusing on the fake tweets column (*FN*), we can see that even for this topic the amount of tweets published by *credulous* users (245k, the 60.85% of FN tweeted by human-operated accounts) is greater than *not-credulous* users (157k, just the 39.15%).
Since the number of FN tweets (402k) is quite similar to the number of RN tweets (326k), this time is useless to do analysis on the 4th and 5th columns, reported just for sake of completeness w.r.t. Table 7.5. By looking at bots, they authored a lot of real tweets, precisely 468,907 (*RN*), that represents more than the 50% of all real tweets; conversely, they published only 116,398 fake tweets (FN), less than the 20% of *Tot.*–*FN*. In our opinion, this is the reason because the amount of *not-credulous* users' real tweets does not overcome the number of the fake ones (2nd column). However, we want just to emphasise the threat of the fake news, and even in this case the *credulous* users "win" for number of fake tweets by a margin of 20 percentage points (higher in the political topic, where the margin is of 39%).

The last perspective we consider is related to observe the number of credulous users that tweeted fake news w.r.t not credulous users. It is important to stress that with this kind of analysis we aim to see if the number of credulous authors spreading fake news is higher than the number of not credulous users in the same task.

Tables 7.7 and 7.8 present the results by looking to what extent, for each topic, the three categories of users (macro-columns' headers in both tables), are participating.

	Credulous			**NotCredulous**		
	#Users	Average	St.Dev.	#Users	Average	St.Dev.
#Fake News≥ 1	54,828	1.54	2.79	19,525 $\downarrow_C$	1.90	5.05
#Real News≥ 1	138,113	1.43	2.91	57,839	1.98	10.22
Only Fake News	47,083	1.40	2.15	13,526 $\downarrow_C$	1.60	4.61
Only Real News	130,368	1.37	2.53	51,480	1.78	10.25

	Bot		
	#Users	Average	St.Dev.
#Fake News≥ 1	9,622	2.07	4.53
#Real News≥ 1	45,924	2.36	9.11
Only Fake News	7,658	1.79	3.78
Only Real News	17,515	2.15	8.80

TABLE 7.7: Users that tweeted in political topic

	Credulous			**NotCredulous**		
	#Users	Average	St.Dev.	#Users	Average	St.Dev.
#Fake News≥ 1	147,158	1.66	8.01	56,451 $\downarrow_C$	2.79	33.56
#Real News≥ 1	39,528	5.30	143.59	19,047	6.10	88.26
Only Fake News	139,187	1.45	6.97	49,351 $\downarrow_C$	1.81	6.51
Only Real News	31,557	1.35	2.96	11,947	1.52	4.72

	Bot		
	#Users	Average	St.Dev.
#Fake News≥ 1	25.818	4.51	22.00
#Real News≥ 1	14,620	33.30	606.39
Only Fake News	19,540	2.39	12.79
Only Real News	8,342	2.19	10.95

TABLE 7.8: Users that tweeted in gossip topic

Specifically, it is reported the amount of users that are authors of: at least a fake tweet (1st row, *#Fake News*$\geq$ 1), at least a real tweet (2nd row, *#Real News*$\geq$ 1), only fake tweets (3rd row, *#Only Fake News*) and only real tweets (4th row, *#Only Real News*). For each of these four cases, the average and standard deviation have been calculated in both tables to show the fake/real tweet's rate and the uniformity of the users belonging to each of the aforementioned cases.

The results reported in Table 7.7 are referred to the topic of political news. The amount of *credulous* users tweeting at least a fake news (1st row) is 54,828, with a publishing rate of 1.54 (tweets per user) on average, and a standard deviation of 2.79. Moreover, we found that 138,113 *credulous* users published at least a real news (2nd row); and despite they are more numerous than previous case, their related tweeting rate (on *average*) is slightly lower (1.43) with an higher standard deviation (2.91).
As regards the amount of *credulous* users publishing *only* fake/real news (3rd and 4th line), we can observe a small numerical decrease in quantity. There are 47,083 *credulous* authors with a publishing rate of 1.40 tweets fake news on average and the standard deviation is of 2.15. In the other case (4th row), 130,368 *credulous* users posted only tweets of real news showing almost the same average (1.37) as in its dual case (only fake news) but with an higher standard deviation (2.53).
About *not-credulous* users, we can notice that the authors posting at least a tweet containing a real news (2nd row) are 57,839, so almost 3 times more than the quantity of not credulous users that at least tweeted a fake news (1st row), i.e., 19,525 (highlighted with $\downarrow_C$ in Table 7.7, w.r.t. the credulous users' value). The averages are similar in both cases, 1.98 for *Real News* and 1.90 for *Fake News*, but the respective standard deviations assume very high values, 5.05 (1st row) and 10.22 (2nd row). This makes us suspicious about the presence of a group of users that tweets a lot. By observing those authors tweeting only fake/real news (3rd and 4th line), we can observe a similar trend to the same case of *credulous* users, but with a much lower level of *participation*. In fact, despite the *not-credulous* authors posting only tweets of real news (51,480) are more than the ones publishing only fake tweets (13,526), these latter authors are only a third of the number of credulous users who only publish fake news. Furthermore, the *not-credulous* users' tweeting rate of real news (*average*) is also higher than the one referred to the ones posting only fake news.
For sake of comparison to human accounts, in Table 7.7 we report the results related to *bots*. We can see a certain disparity between the number of bots tweeting *Fake News* (i.e, 9,622 as indicated in 1st row) and *Real News* (i.e., 45,924 as indicated in 2nd row). On the other hand, by comparing the data of this case (i.e., at least a fake/real tweet) with the ones related to bots authoring *only* fake/real news (3rd and 4t line), we can see: (i) a reduction equal to more than 2 times about bots tweeting *only* real news (17,515 in

4th line w.r.t the case in the 2nd line), and (ii) a little reduction for what concerns the amount of bots sharing only fake news in their tweets (7,658 in 3rd line, w.r.t. 1st line). The tweeting averages per bots are upper than the values corresponding both human's cases (i.e., credulous and not credulous users), but not so much; the standard deviations have higher values when referred to real news (9.11) and *only real news* (8.80).

The outcomes concerning to gossip topic are presented in Table 7.8. Starting by the 1st macro-column (headed, *credulous*), we can see a big amount of authors having at least one fake-news tweet (147,158), and this numerical superiority (w.r.t. the number of credulous users tweeted real news) occurs even when we count the ones tweeting *only* fake news (139,187). Indeed, about the authors of real news' tweets, 39,528 credulous users have at least one and 31,557 published *only* real tweets.
By looking at the details of not credulous users, we can see a good downsizing of fake news' authors (highlighted with $\downarrow_C$ in Table 7.8, w.r.t. the credulous users' value). Precisely, there are 56,451 users that tweeted at least one fake news, and 49,351 users that published only fake news in their tweets. Moreover, we observed the same decreasing trends also in both cases of real news published by *not-credulous* authors. In particular, the authors publishing at least one real news are 19,047 (2nd row), whereas the ones publishing only real news are 11,947.
About the automated accounts (*bot*, 3rd macro-column), the tweeting bots of at least a fake news are 25,818 and those ones publishing only fake are 19,540; more than the ones publishing real news which are 14,620 (2nd row) and 8,342 (4th row).

It is worth to stress that in Table 7.8, with regard to the average and standard deviation, the values are very high compared to the other lines, especially focusing on the values reported in the 2nd row. As already suspected in the previous case of political news, here we are almost certain of the presence of a very small group of users who tweet compulsively and that are responsible of such strange values, but at least they are publishing just true news (hence harmless). However, this little drawback does not overshadow the fact that credulous users, compared to the not credulous ones, are more involved in posting fake news.

Overall, the averages of the bots is higher than both *credulous* and *not-credulous* users, regardless the news' veracity in their tweets. The fact that the statistical descriptors' values (of fake and real news) are similar and not too much higher than the values related to humans, may mean some bots strategy to avoid detection, i.e., by mimicking, in such a way, human's behaviour.

7.4 Discussion

The experimental results described in this chapter shed light on the connection between fake news and *credulous* users, and on the extent they are involved in fake news spreading on Twitter. We investigated such connection under three perspectives in order to have a wider view. The best and the most important findings come from the *tweeting* perspective. We found that the tweets authored by *credulous* users are always more (in quantity) then those published by *not credulous* accounts. This was expected about fake news; however, it was unexpected for the real news. Fake news actually represent the problem, real one are harmless, and it would be a bit extreme to expect that *credulous* users are active exclusively on fake news. Moreover, about fake news, we discovered a considerable involvement of credulous users in spreading tweets containing fake news (especially about politics). From the *news* perspective, it seems that credulous users are less interested in fake gossips (only 80% of fake news) with respect to fake politics (95%). In the last perspective, we had some difficulties to evaluate the participation of users because of uncommon posting activity of some users. As far as fake tweets are concerned, more credulous users participate than the not credulous ones.

Inspired by these encouraging results, we are able to give an answer also to the following research question.

> **RQ5** – Do credulous users contribute to fake news spreading? What is their level of involvement compared to that of not credulous users and bots? Is it possible to provide evidence of credulous users contributing to misinformation? Can we take advantage of credulous users detection for fake news detection?

ANSWER – *Mis-/dis- information is driven by the production and circulation of fake news, and we can say that credulous users are very active entities on OSM in spreading them (see Tables 7.5 and 7.6).*
Observing the high percentage of fake news tweeted by credulous users (61% in politics and 70% in gossip), we can point out the harmfulness they constitute for their followers. To have an idea about how to benefit from credulous detection, let us assume to suspend or remove (extreme case) credulous users' accounts from Twitter; in this case, the 61%-70% of fake news produced by human-operated accounts would automatically disappear.

This first evidence that credulous users (and not only bots) are involved, even if unconsciously, in supporting malicious activities (spreading disinformation) can inspire further research in this direction. For instance, by using tweets' provenance to enhance the effectiveness of fake news detector, tweets from credulous users should have a higher probability to be fake news. In such a way, it would be possible also to study the credibility [168] of credulous users as a source of information.

Chapter 8

Conclusion

As highlighted by the Reuters Institute [123] and recent literature [132, 166, 178, 189], social media are gradually becoming a very effective means of information mainly used by people to keep up with news, especially by the youngsters. There are two main factors that characterise OSM's effectiveness in communication: (i) faster news dissemination, and (ii) capability to reach large audiences, with respect to traditional mass media such as television, newspapers and radio. In fact, domestic users of such services can keep up with the news effortlessly, while routinely checking out their social channels of interest. Unfortunately the lack of effective methods for the automatic fact-checking, together with an impressive rate of posting on social platform, favoured the rise up of mis-/dis- information phenomena through the increasingly use of fake news; hence, exposing OSM's users at an important risk of being misinformed. Mis-/Dis- information becomes even more effective when it is targeted towards certain categories of users (*targeted misinformation*). Since the Facebook-Cambridge Analytica scandal, several other events start to attract the attention of academicians due to the suspect of external intervention to bias the results, like Brexit (in 2016), US Presidential Election (in 2016), Kenya elections (in 2013 and 2017) and many others[1].

Academics, governments and OSM administrators agree that responsible of the production and proliferation of fake news are (*social*) *bots*. They are (totally or partially) automated accounts in OSM that, by acting under fictitious identities and mimicking human's behaviour on social networks, actively interact with genuine human users to capture their interests. The goal is to pursue malicious purposes like: hate speeches, generate discontent and misconception and, in general, to induce a bias within their opinion to affect their mood.

[1] https://tinyurl.com/yxlo8f3u

The role of bots is to influence human users by means of misinformation and fake news. In [51] it has been highlighted that the human users do not pay attention when sharing and publishing news. As consequence, although sometimes unknowingly, such users contribute to mis-/dis-information activities of Social Media malicious entities.

Misinformation may largely affect the opinion of human users in several (real-world) domains, along with the malicious activities continuously carried on by malevolent social media entities (e.g.., bots). In this thesis, by using Twitter as a benchmark, we deliberately draw attention to those human users who are particularly exposed to fake news attack (targeted misinformation attacks) performed by (net of) bots.

In Section 1 we formulated five research questions that were driving our research's investigations. The experiments and analyses (see Chapters 3–7) provided some answers to shed the light on the effectiveness of misinformation when conducted in a targeted manner, and on a particular category of human users.

In Chapter 3, we answered to the first research question reported hereafter.

> **RQ1** – Among human Twitter users, which type of social relationship (e.g., following or being followed) is the most influential, and why? Does it make sense to assign a *gullibility* score to human users? Which user-related aspects should be taken into account in such a score? Does a clear separation between credulous and not credulous users exist? Or, simply, is one user more credulous than another?

In that chapter, by abuse of language, we have provided the definition of *credulous* users as all those human-operated accounts on Twitter who follow a high percentage of bots (*bot-followees*, which means being follower of a bot) with respect to total number of their *followees*. Trivially, following a high number of bots may increase the probability to see on the own dashboard the content (potentially deceptive) they publish. To identify the credulous users we set up a method, based on four heuristics, to rank human users by analysing the nature (human or bot) of their followees. The effectiveness of this ranking method has been tested by means of an *efficacy* measure, i.e., the ratio between the bot-followees of a group of users (single out as credulous) over the total number of bots in our dataset. We found that on average credulous users have a bot-followees concentration ranging from 30% to 36% that, compared to what stated in [61] (i.e., bot accounts in Twitter range from 8 to 15%), is around the double. By using this method, we provide a first dataset of credulous users that constitutes a preliminary ground truth of such human-operated accounts. However, the limitation of this approach is that it is rather expensive due to the need of retrieving information of many (potentially) followees.

In Chapter 4 we tried to overcome this limitation by using ML techniques. Decision models were built to recognise *credulous* users avoiding the followees inspection. Below we recall the second research question we addressed.

> **RQ2** – How effectively Machine Learning (ML) techniques can be in distinguishing credulous and non-credulous users? Is it possible to avoid in depth inspection of human users' social contacts in order to lighten the complexity of identifying credulous users? What is the loss in terms of accuracy when performing their identification? What are the features of Twitter accounts that can facilitate this distinction? Are the features used for bot detection beneficial also for identifying credulous users?

To answer these questions, we built a larger ground-truth constituted by 316 credulous and 2,522 not credulous users. By means of an extensive experimentation (with 19 learning algorithms), we obtained a credulous classifier that (on average) achieve the 93.27% of instances correctly classified (*accuracy*) and an AUC of 0.93. When looking at the usefulness of features, users' profile data only resulted to be the most relevant. We also investigated the effectiveness of features designed for bot detection but we obtained bad classification results, showing they are uselessness for this task. We found also some very useful features that emphasise the difference between credulous and not credulous users; i.e.,: #friends/#followers^2, #friends, following rate, 2 x #followers $\geq$ #friends, and #followers. Despite it was not our primary target, we also obtained very good classification performance in bot detection, i.e., an *accuracy* of 98.41% (AUC 1.00) that is slightly better than values of various bot detectors reported in [42].

In Chapter 5 we tried to predict the percentage of bots a human user is following. The motivation for such an investigation is to find a methodology able to identify, in a purely preventive way, human users who may end up in the target of misinformation attacks by bots. This investigation contributed to answer to our third research question.

> **RQ3** – Is it possible to predict the number of bots a human user is following (*bot-followees*)? Are the features, used for credulous classification, useful also for this task? Which measures can be adopted to estimate the quality of such predictions in absence of well-defined benchmarks in the literature?

By adopting the same features used for credulous classification task, we have not obtained good results. To improve the regression performance, we considered the union of two feature sets, i.e., the ones designed for bot detection, in addition to the ones used for credulous detection. This way, we obtained an acceptable margin of error (Mean Absolute Error) when predicting the bot-followees percentage a human-operated account

is following. Since [149, 169] we can confirm it is a challenging task and, despite we got a regression model performing significantly better than the baseline (provided by the *ZeroR*[2]), it would be nice to compare our results with the literature. At best of our knowledge there is no a clearly defined baseline.

In Chapter 3 and 4 (and marginally in 5), we designed strategies to identify credulous users, but balancing between need of data and computational costs. In Chapter 6 we started to investigate the behaviour of credulous users with the goal to understand if they actively participate in bouncing content from bots and therefore potentially malevolent. The fourth research question helped to target our analysis by focusing on users' behavioural aspects.

> **RQ4** – Is it enough to compare the different types of activities (i.e., retweets, quoted tweets, replies and posting new content) between credulous and not credulous users to significantly differentiate them? Can bot-followees influence, in terms of content production, the activities of credulous users more than not credulous ones? How to measure the effectiveness of such an influence? Do credulous users bounce bots' content? And to what extent with respect to not credulous users?

To find differences between the two categories of users, we conducted a coarse-grained analysis in which the activities (i.e., retweets, replies and handmade tweets) of credulous and not credulous users have been compared in statistical terms. Excluding the case of handmade tweets (called pure tweets), where we found a very different tweet-production rate, by observing the statistics of the other types of tweets (i.e., replies and retweets) no significant behavioural differences emerged. Such a behavioural similarity (in this coarse-grained analysis), between the two populations, is mainly due to the human nature that both credulous and not credulous users share. Indeed, behavioural differences concerning the activities on OSM are more marked/evident when human and bot accounts are compared; and often adopted to train bot detectors [45].
Conversely, by performing a fine-grained analysis we got relevant differences between credulous and not credulous users. Precisely, we found that credulous users bounce more bot-originated content than not credulous ones on average. Such differences have been tested by performing non-parametric statistical test (e.g., ANOVA and Mann-Whitney), revealing a significant difference between the two populations.
The findings obtained from the fine-grained analysis allow us to claim that some relevant differences between credulous and not credulous users exist. In particular, from the analysis carried out on each type of tweet, we can state that the credulous users' tweets

[2]ZeroR: a classifier that predicts the mean in case of numeric/real class values.

production is significantly influenced by bots activities on OSM. In fact, a considerable percentage of contents posted (retweeted, quoted or replied) by credulous users (unlike not credulous users) come from automated (and potentially malicious) accounts.

These findings confirm our suspect about the involvement of credulous users in supporting bots activities in terms of content dissemination (potentially malicious). Our credulous classifier is helpful in single out those human-operated accounts active in this business. In Chapter 7 we shed light on the harmfulness of content spread by credulous users. The fifth research questions has been addressed by the analysis conducted on Chapter 7.

> **RQ5** – Do credulous users contribute to fake news spreading? What is their level of involvement compared to that of not credulous users and bots? Is it possible to provide evidence of credulous users contributing to misinformation? Can we take advantage of credulous users detection for fake news detection?

Starting from a publicly available fake news dataset (*FakeNewNet*) where, for each news, a list of tweets ids containing that news was provided, we traced back to the Twitter accounts posting that news. We firstly applied our bot detector, to filter out the bots, and then, on the remaining accounts, we applied our approach for credulous detection. We found that a large percentage of tweets containing fake news have been posted by credulous users (61%-70% of the fake news diffused by human-operated accounts). This result provides evidence of the level of involvement that credulous users have in misinformation activities, w.r.t. not credulous ones, on Social Media.
As far as bots are concerned, we detected few automated accounts (bots) if compared with the number of human users. Therefore, we repute the number of tweets authored by these bots not so high to perform a fair comparison with the categories of human users (i.e., credulous and not credulous users). Anyway, on average, we noticed that bots' tweeting rate is larger than the humans' one, even if not too much. Furthermore, we observed that the bots tweeting rate is similar between fake and real news. We suspect that both the similarity with the humans' posting rate (thus mimicking human behaviour [106]) and the balancing of fake/real news rate are part of a strategy to avoid their (bot) detection on Social Media. Another interesting detail emerged from our analysis of bots concerns the percentage of automated accounts, out of the total number of users we considered. The percentage of bots we detected, by means of our bot detector, is of 10.55%, supporting what stated in [61] about the percentage of bots active on Twitter (ranging from 8-15%).

The analysis conducted in this thesis shows a active humans' participation in spreading fake news, demonstrating the strong involvement of credulous users in mis-/dis-

information dissemination. There are several benefits on identifying credulous users, for instance, they can be considered as natural honeypot to attract/study and remove bots. A possible application can be the inspection of the data stream published by *credulous* users', with targeted fact-checking. This way, it is reduced the set of human users to scrutiny and, indirectly, the number of tweets, to perform targeted fact-checking. To narrow the group of *credulous* users to analyse, it might be helpful to implement priority measures that consider the content production's rate of *credulous* users (to pay more attention on the most active ones), or the number of followers (i.e., other potential victims). OSM administrators represent one of the stakeholders of such systems interested to credulous users' activities (e.g., content re-posting). They can slow down the propagation of malicious information, contributing not only in fake news fighting, but also in improving the credibility and effectiveness of their platforms as a mass media.

8.1 Future research directions

Despite these encouraging and promising results, additional efforts are needed to improve users' awareness in news trustworthiness. In the following we propose some interesting future research directions:

(i): observe the variations of credulous users' followees and check, by considering an observation time frame, the (human/bot) nature of those started of being followed (new followees), those stopped of being followed, and those remaining longer in their followees lists. Also looking at the amount of changes in profile data (in a given time frame) can be interesting, recalling their employment as the most effective features for credulous detector are calculated (*ClassA*-'s features). Thereby, we can investigate their ability to detect potentially malicious accounts and derive further features for credulous classification;

(ii): initially started in Chapter 5, it would be proficient to develop approaches for credulous detection also to those human-operated accounts with more than 400 followees. Investigations in this direction would further contribute to shed light on users' influence mechanisms; for example, to understanding whether the proportion of suspicious users that a credulous user follows is proportional to the number of followees.

(iii): initially started in Chapter 7 through the consideration of two topics (politics and gossip), it would be interesting to relate the topic of (fake) news to the level of involvement of credulous users considering, for example, more current and topical news than gossip, such as healthcare and science;

(iv): since the concept of credulous users is strongly connected to the type of relationship between users on a specific social platform, it would be interesting to produce or at least adapt the concept of credulous users for other OSM platforms (e.g. Facebook). Depending on the type of social relationships that the platform offers, the concept of being interested in content published by an account also changes; this makes the latter direction of research very challenging, maybe the most one.

Appendix A

Automatic Detection of Credulous Users on Twitter – Complete Results

The adopted machine learning algorithms can be grouped in five categories. The first category is called *bayes* and it includes: Hidden Markov Models (HMM) [54], Bayesian Networks (BN) [65], and Naive Bayes (NB) [85]. The second category is named *lazy*, and we only use the K-nearest neighbours (IBk) [2]. The first category, namely *functions*, includes: Neural Networks with back propagation of error to learn a multi-layer perceptron (MLP) [127], Voted Perceptron [63] (VP) and Sequential Minimal Optimization (SMO)[136]. *Rules* is the fourth category including: RIPPER (JRip) [37], 0R and 1R [79]. Finally, there is a category called *trees* that comprises: C4.5 (J48) [138], Hoeffding tree (HT) [83], Random Decision Tree (RT) [77], Consolidated Tree Construction (J48c) [133], Grafted C4.5 Decision Tree (J48g) [175], a decision tree builder using the LogitBoost strategy (LAD) [78], Logistic Model Tree (LMT) [93], another one using information gain/variance and Reduced-Error Pruning with backfitting (REP) [137] and, Random Forest (RF) [24].

A.1 Bot Detection - complete results

Table A.1 reports the complete results of the experimentation that has been performed for bot detection.

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	HMM	55.28	0.55	1.00	0.71	0.50
	IBk	97.34	0.97	0.98	0.98	0.97
	BN	96.98	0.98	0.97	0.97	0.99
	NB	97.03	0.98	0.97	0.97	0.98
	VP	81.24	0.83	0.83	0.83	0.81
	MLP	97.91	0.98	0.98	0.98	0.99
	SMO	98.04	0.98	0.98	0.98	0.98
	JRip	97.92	0.99	0.98	0.98	0.99
	1R	95.29	0.96	0.96	0.96	0.95
	0R	55.28	0.55	1.00	0.71	0.50
	J48	97.75	0.98	0.98	0.98	0.98
	HT	96.66	0.97	0.97	0.97	0.97
	RT	96.85	0.97	0.97	0.97	0.97
	J48c	97.78	0.98	0.97	0.98	0.98
	J48g	97.88	0.98	0.98	0.98	0.98
	LAD	97.44	0.98	0.98	0.98	0.99
	LMT	98.15	0.99	0.98	0.98	0.99
	REP	97.67	0.98	0.98	0.98	0.99
	RF	**98.33**	**0.99**	**0.98**	**0.98**	**1.00**
Botometer+	HMM	55.28	0.55	1.00	0.71	0.50
	IBk	97.05	0.97	0.97	0.97	0.97
	BN	96.99	0.97	0.97	0.97	0.99
	NB	97.17	0.98	0.97	0.97	0.99
	VP	93.52	0.97	0.91	0.94	0.94
	MLP	97.78	0.98	0.98	0.98	0.99
	SMO	97.64	0.98	0.98	0.98	0.98
	JRip	97.61	0.98	0.97	0.98	0.98
	1R	95.25	0.96	0.96	0.96	0.95
	0R	55.28	0.55	1.00	0.71	0.50
	J48	97.53	0.98	0.97	0.98	0.98
	HT	96.72	0.97	0.97	0.97	0.97
	RT	96.52	0.97	0.97	0.97	0.96
	J48c	97.49	0.98	0.97	0.98	0.98
	J48g	97.60	0.98	0.97	0.98	0.98
	LAD	97.32	0.98	0.97	0.98	0.99
	LMT	97.79	0.98	0.98	0.98	1.00
	REP	97.45	0.98	0.97	0.98	0.99
	RF	**97.97**	**0.98**	**0.98**	**0.98**	**1.00**
ClassA-	HMM	55.28	0.55	1.00	0.71	0.50
	IBk	91.03	0.91	0.93	0.92	0.91
	BN	87.15	0.93	0.83	0.88	0.94
	NB	64.37	0.89	0.42	0.54	0.77
	VP	80.07	0.82	0.82	0.82	0.80

	MLP	85.01	0.89	0.84	0.86	0.91
	SMO	68.58	0.76	0.63	0.69	0.69
	JRip	94.38	0.96	0.94	0.95	0.96
	1R	84.51	0.88	0.84	0.86	0.85
	0R	55.28	0.55	1.00	0.71	0.50
	J48	94.30	0.96	0.94	0.95	0.96
ClassA-	HT	84.48	0.90	0.81	0.85	0.88
	RT	92.48	0.93	0.94	0.93	0.92
	J48c	94.36	0.96	0.93	0.95	0.96
	J48g	94.41	0.96	0.94	0.95	0.96
	LAD	89.19	0.93	0.87	0.90	0.94
	REP	93.96	0.96	0.93	0.94	0.97
	LMT	94.33	0.96	0.94	0.95	0.97
	RF	**95.84**	**0.98**	**0.95**	**0.96**	**0.99**

Table A.1: Complete results for bot detection

A.2 Credulous Detection - complete results

A.2.1 Main experiments

Table A.1 reports the complete results of the experimentation that has been performed for credulous users classification (316 Credulous *vs.* 2,522 Not Credulous users).

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
	HMM	50.06	0.50	1.00	0.67	0.50
	IBk	89.69	0.74	0.73	0.90	0.96
	BN	80.26	0.91	0.89	0.76	0.91
	NB	73.41	0.91	0.68	0.73	0.73
	VP	68.68	0.72	0.63	0.67	0.70
	SMO	78.77	0.80	0.78	0.78	0.79
	MLP	77.76	0.79	0.77	0.78	0.85
	JRip	92.80	0.99	0.87	0.92	0.93
	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
ALL_features	0R	49.51	0.49	0.65	0.66	0.50
	J48	91.62	0.94	0.90	0.91	0.94
	HT	80.21	0.91	0.68	0.76	0.90
	RT	86.33	0.87	0.86	0.86	0.86
	J48C	91.73	0.94	0.90	0.92	0.94
	J48g	91.82	0.94	0.90	0.92	0.94
	LAD	92.38	0.95	0.90	0.92	0.97
	LMT	91.63	0.95	0.88	0.91	0.96
	REP	93.07	0.99	0.88	0.93	0.94

	RF	92.16	0.96	0.88	0.92	0.97
Botometer+	HMM	50.06	0.50	1.00	0.67	0.50
	IBk	65.03	0.61	0.60	0.63	0.70
	BN	61.02	0.67	0.62	0.50	0.69
	NB	60.44	0.68	0.42	0.60	0.60
	VP	56.45	0.61	0.64	0.54	0.59
	SMO	64.63	0.68	0.59	0.61	0.65
	MLP	64.72	0.67	0.58	0.61	0.69
	JRip	66.42	0.67	0.67	0.66	0.67
	1R	63.54	0.63	0.65	0.64	0.64
	0R	49.51	0.49	0.65	0.66	0.50
	J48	66.02	0.68	0.63	0.63	0.66
	HT	61.05	0.67	0.47	0.51	0.68
	RT	60.93	0.61	0.61	0.61	0.61
	J48C	65.67	0.67	0.64	0.63	0.66
	J48g	66.16	0.68	0.63	0.63	0.67
	LAD	65.79	0.67	0.65	0.65	0.69
	LMT	67.20	0.69	0.62	0.65	0.72
	REP	65.75	0.66	0.66	0.66	0.68
	RF	**67.81**	**0.68**	**0.69**	**0.68**	**0.73**
ClassA-	HMM	50.06	0.50	1.00	0.67	0.50
	IBk	92.59	0.74	0.73	0.92	0.97
	BN	82.77	0.98	0.88	0.79	0.93
	NB	73.00	0.97	0.69	0.73	0.73
	VP	68.68	0.72	0.63	0.67	0.70
	SMO	75.32	0.74	0.80	0.77	0.75
	MLP	80.08	0.81	0.81	0.80	0.87
	JRip	93.05	0.99	0.87	0.93	0.94
	1R	**93.27**	**0.99**	**0.88**	**0.93**	**0.93**
	0R	49.51	0.49	0.65	0.66	0.50
	J48	92.58	0.97	0.88	0.92	0.94
	HT	83.28	0.96	0.71	0.80	0.93
	RT	88.88	0.89	0.89	0.89	0.89
	J48C	92.68	0.97	0.88	0.92	0.94
	J48g	92.64	0.97	0.88	0.92	0.94
	LAD	92.38	0.96	0.89	0.92	0.97
	LMT	92.66	0.98	0.88	0.92	0.96
	REP	93.09	0.98	0.88	0.93	0.95
	RF	92.71	0.97	0.89	0.92	0.97

TABLE A.2: Complete results for credulous detection – 316 Credulous users

A.2.2 Additional experiment - cut to 946 users

Table A.3 reports the complete results of the experimentation that has been performed for credulous users classification by considering 443 users as credulous.

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
ALL_features	HMM	48.31	0.48	1.00	0.65	0.50
	IBk	86.07	0.72	0.71	0.85	0.94
	BN	79.12	0.86	0.85	0.75	0.89
	NB	72.29	0.87	0.69	0.71	0.72
	VP	67.94	0.70	0.62	0.65	0.69
	SMO	78.48	0.80	0.74	0.77	0.78
	MLP	76.19	0.76	0.75	0.75	0.83
	JRip	**89.92**	**0.97**	**0.82**	**0.88**	**0.91**
	1R	89.42	0.97	0.80	0.88	0.89
	0R	51.42	0.40	0.56	0.53	0.50
	J48	88.04	0.90	0.85	0.87	0.90
	HT	78.19	0.87	0.70	0.75	0.89
	RT	83.33	0.83	0.83	0.83	0.83
	J48C	88.23	0.90	0.85	0.87	0.91
	J48g	88.44	0.91	0.85	0.88	0.90
	LAD	89.01	0.92	0.84	0.88	0.95
	LMT	88.83	0.93	0.83	0.88	0.94
	REP	89.79	0.96	0.82	0.88	0.92
	RF	89.58	0.94	0.84	0.89	0.95
Botometer+	HMM	48.31	0.48	1.00	0.65	0.50
	IBk	64.54	0.59	0.58	0.60	0.69
	BN	61.23	0.65	0.57	0.53	0.68
	NB	60.15	0.66	0.48	0.58	0.60
	VP	58.63	0.64	0.53	0.45	0.59
	SMO	64.34	0.67	0.57	0.59	0.64
	MLP	64.84	0.66	0.56	0.60	0.69
	JRip	66.16	0.65	0.64	0.64	0.66
	1R	64.59	0.64	0.62	0.63	0.64
	0R	51.42	0.40	0.56	0.53	0.50
	J48	65.74	0.65	0.62	0.63	0.66
	HT	61.07	0.66	0.45	0.52	0.66
	RT	59.49	0.58	0.58	0.58	0.59
	J48C	64.77	0.63	0.64	0.63	0.65
	J48g	65.85	0.66	0.62	0.63	0.66
	LAD	65.37	0.65	0.62	0.63	0.69
	LMT	**66.76**	**0.68**	**0.59**	**0.62**	**0.71**
	REP	64.92	0.65	0.63	0.63	0.66
	RF	66.36	0.66	0.64	0.65	0.71
ClassA-	HMM	48.31	0.48	1.00	0.65	0.50
	IBk	88.98	0.72	0.70	0.88	0.94
	BN	81.73	0.95	0.82	0.77	0.91
	NB	72.35	0.94	0.70	0.71	0.72

	VP	67.88	0.69	0.62	0.65	0.69
	SMO	73.73	0.73	0.77	0.74	0.74
	MLP	80.31	0.81	0.80	0.80	0.87
	JRip	**90.08**	**0.98**	**0.81**	**0.88**	**0.91**
	1R	89.42	0.97	0.80	0.88	0.89
	0R	51.42	0.40	0.56	0.53	0.50
	J48	89.44	0.94	0.84	0.88	0.92
ClassA-	HT	80.55	0.92	0.72	0.77	0.91
	RT	85.36	0.85	0.85	0.85	0.85
	J48C	89.40	0.94	0.84	0.88	0.92
	J48g	89.49	0.94	0.84	0.88	0.92
	LAD	89.69	0.94	0.84	0.89	0.95
	LMT	89.39	0.95	0.83	0.88	0.94
	REP	89.91	0.96	0.82	0.89	0.92
	RF	89.65	0.94	0.84	0.89	0.95

TABLE A.3: Complete results for credulous detection – 443 Credulous users (*cut946*)

A.2.3 Additional experiment - cut to 1030 users

Table A.4 reports the complete results of the experimentation that has been performed for credulous users classification by considering 502 users as credulous.

		evaluation metrics				
	alg	*accuracy*	*precision*	*recall*	*F1*	*AUC*
	HMM	46.92	0.47	1.00	0.64	0.50
	IBk	85.23	0.71	0.70	0.84	0.93
	BN	78.24	0.84	0.84	0.75	0.89
	NB	72.87	0.83	0.73	0.70	0.72
	VP	68.60	0.70	0.59	0.64	0.69
	SMO	78.28	0.79	0.72	0.75	0.78
	MLP	76.95	0.76	0.74	0.75	0.84
	JRip	**88.45**	**0.95**	**0.79**	**0.86**	**0.89**
	1R	87.49	0.96	0.76	0.84	0.87
ALL_features	0R	52.93	0.38	0.60	0.50	0.50
	J48	86.11	0.87	0.83	0.84	0.89
	HT	75.58	0.83	0.66	0.72	0.85
	RT	82.43	0.81	0.81	0.81	0.82
	J48C	86.28	0.87	0.83	0.85	0.89
	J48g	86.50	0.87	0.83	0.85	0.89
	LAD	87.84	0.91	0.81	0.86	0.94
	LMT	87.49	0.91	0.80	0.85	0.93
	REP	87.99	0.94	0.79	0.85	0.91
	RF	88.31	0.92	0.81	0.86	0.95

	HMM	46.92	0.47	1.00	0.64	0.50
	IBk	66.31	0.57	0.57	0.61	0.70
	BN	60.87	0.65	0.58	0.55	0.68
	NB	60.00	0.64	0.54	0.57	0.59
	VP	60.43	0.56	0.55	0.46	0.60
	SMO	64.81	0.63	0.46	0.47	0.62
	MLP	65.16	0.64	0.57	0.59	0.69
	JRip	67.27	0.65	0.62	0.63	0.67
	1R	67.14	0.65	0.61	0.63	0.66
Botometer+	0R	52.93	0.38	0.60	0.50	0.50
	J48	66.16	0.64	0.60	0.61	0.65
	HT	62.66	0.63	0.38	0.44	0.65
	RT	59.33	0.56	0.57	0.56	0.59
	J48C	64.34	0.61	0.64	0.62	0.64
	J48g	66.32	0.65	0.60	0.61	0.65
	LAD	66.33	0.64	0.61	0.62	0.69
	LMT	**67.70**	**0.68**	**0.58**	**0.62**	**0.71**
	REP	66.05	0.64	0.60	0.61	0.67
	RF	66.87	0.65	0.62	0.63	0.71
	HMM	46.92	0.47	1.00	0.64	0.50
	IBk	87.25	0.71	0.70	0.85	0.93
	BN	82.99	0.92	0.81	0.80	0.91
	NB	72.76	0.90	0.76	0.70	0.72
	VP	68.54	0.70	0.59	0.63	0.69
	SMO	73.58	0.72	0.76	0.73	0.73
	MLP	80.94	0.82	0.78	0.79	0.88
	JRip	**88.70**	**0.96**	**0.79**	**0.86**	**0.89**
	1R	87.49	0.96	0.76	0.84	0.87
ClassA-	0R	52.93	0.38	0.60	0.50	0.50
	J48	87.67	0.92	0.81	0.85	0.91
	HT	78.76	0.88	0.72	0.75	0.89
	RT	83.74	0.82	0.82	0.82	0.83
	J48C	87.54	0.91	0.81	0.86	0.90
	J48g	87.72	0.92	0.81	0.85	0.91
	LAD	88.32	0.94	0.80	0.86	0.94
	LMT	88.03	0.94	0.79	0.86	0.93
	REP	88.39	0.94	0.80	0.86	0.92
	RF	87.86	0.92	0.81	0.86	0.94

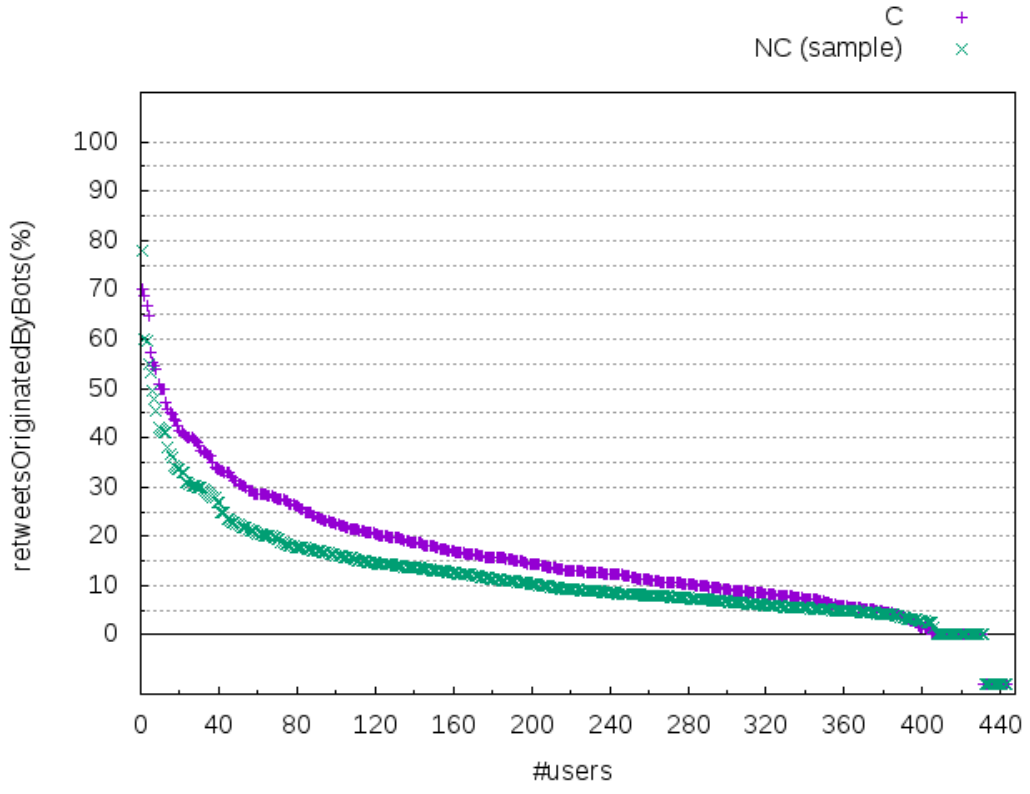
TABLE A.4: Complete results for credulous detection – 502 Credulous users (*cut1030*)

Appendix B

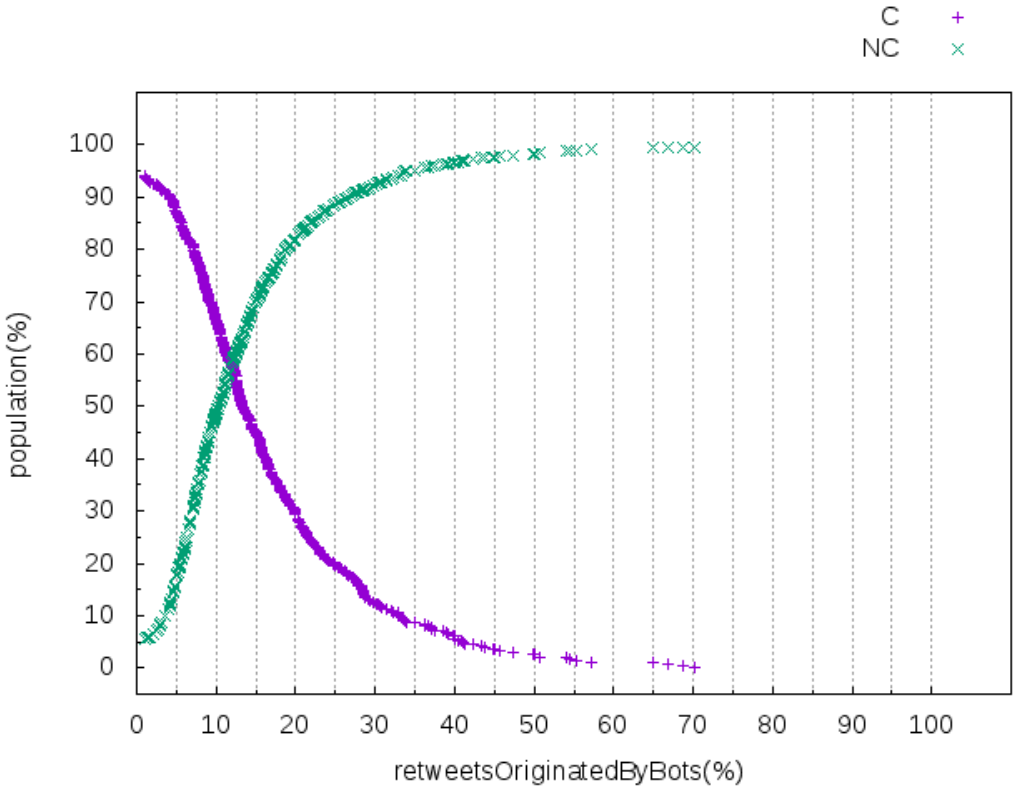
Behavioural Analysis: Extended Investigation Results

B.1 *cut946*

B.1.1 Retweets

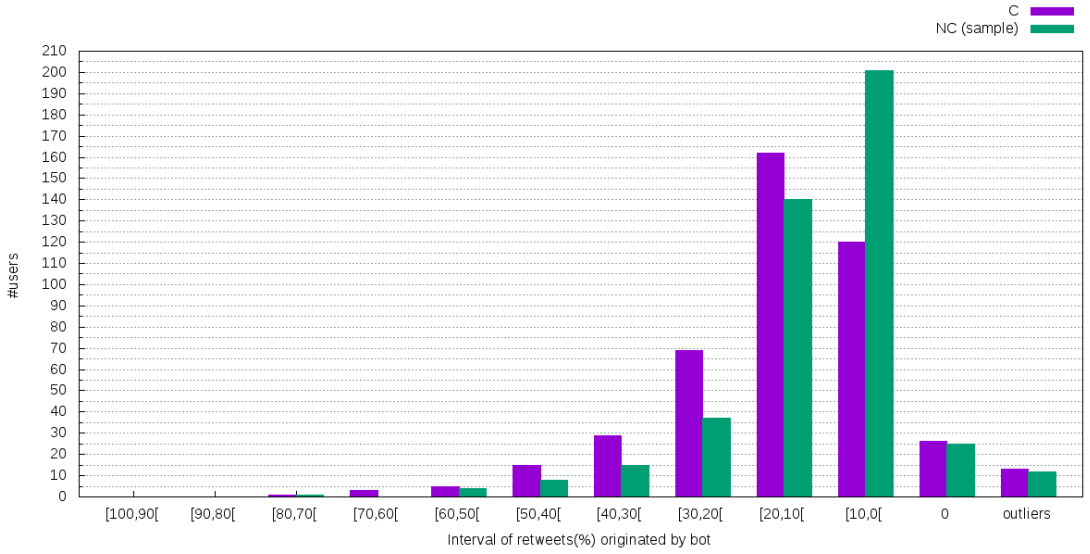


(A) Percentage of 'byBots'-retweets posted by C and NC (sample) users.

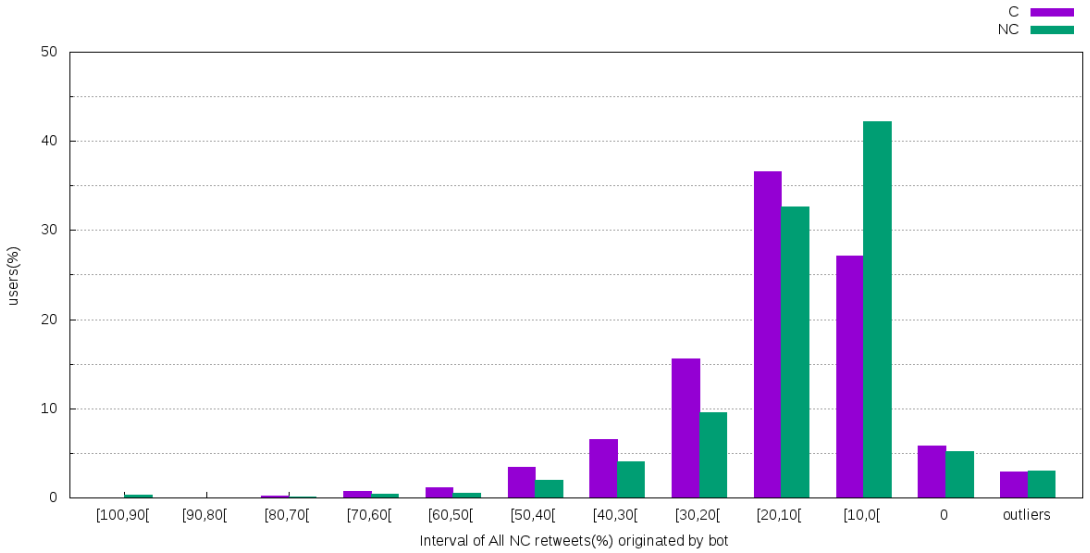


(B) % of populations w.r.t. the % of 'byBots'-retweets.

FIGURE B.1: Comparative analysis between C and NC users w.r.t. 'byBots'-retweets. Here, the set of C users includes 443 humans (namely, *cut946*).



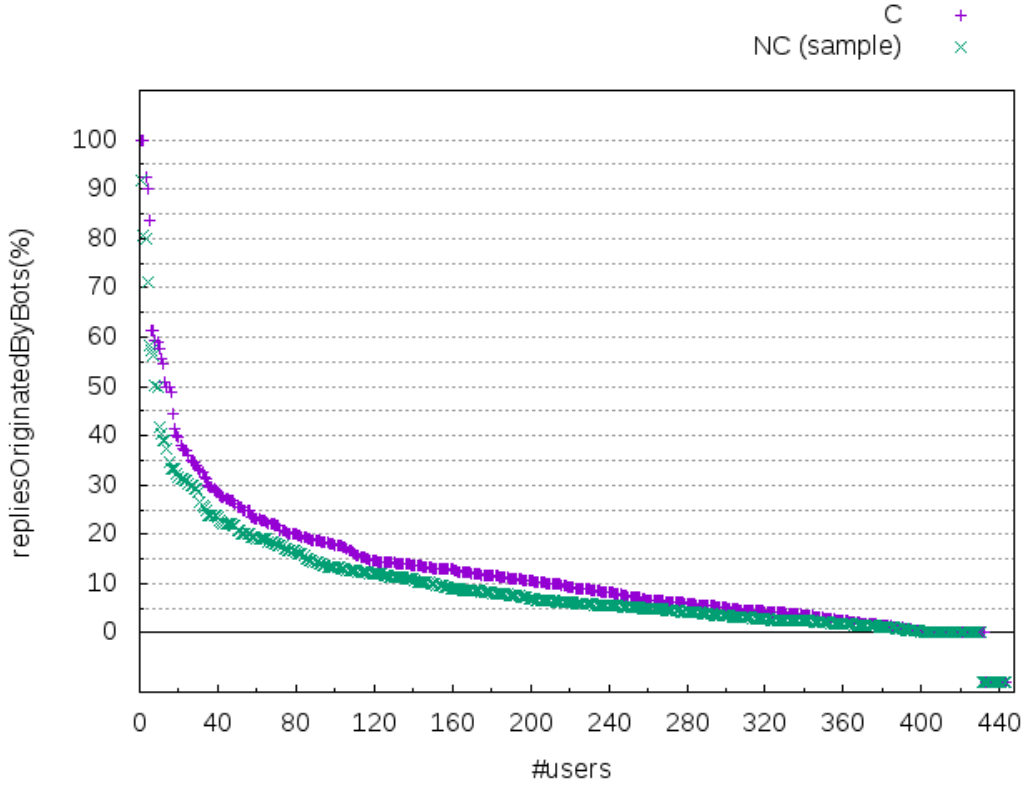
(A) Deciles of Figure B.1a.



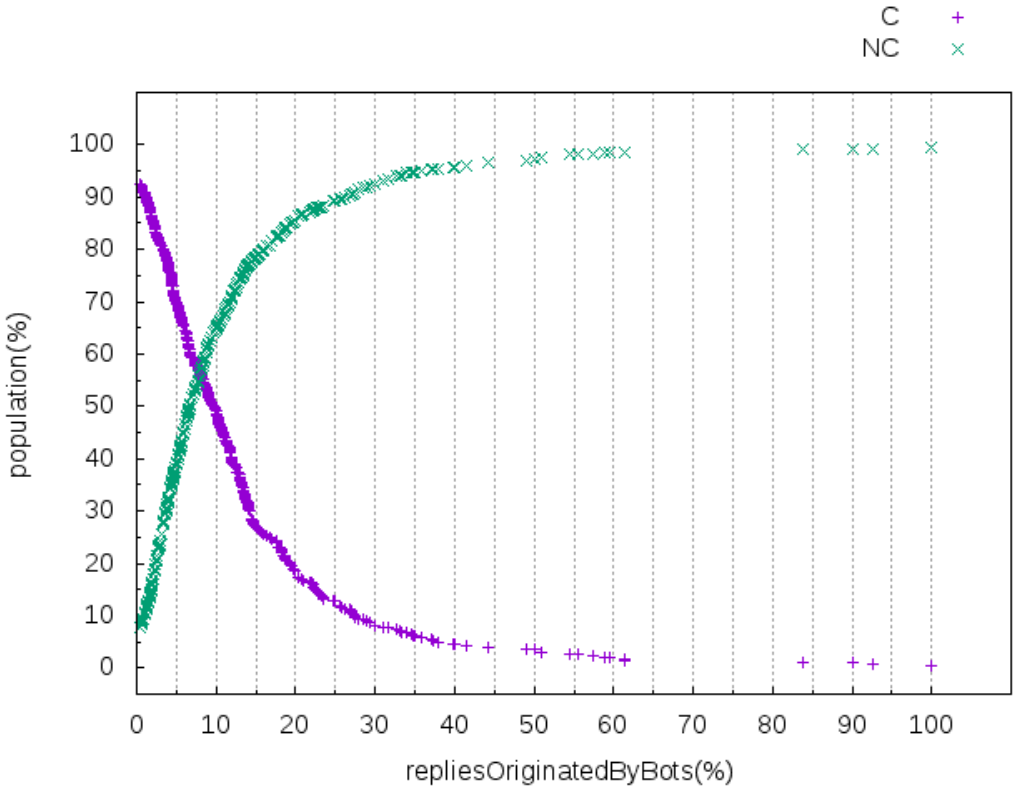
(B) Deciles of C and all NC users.

FIGURE B.2: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-retweets. Here, the set of C users includes 443 humans (namely, *cut946*).

B.1.2 Replies

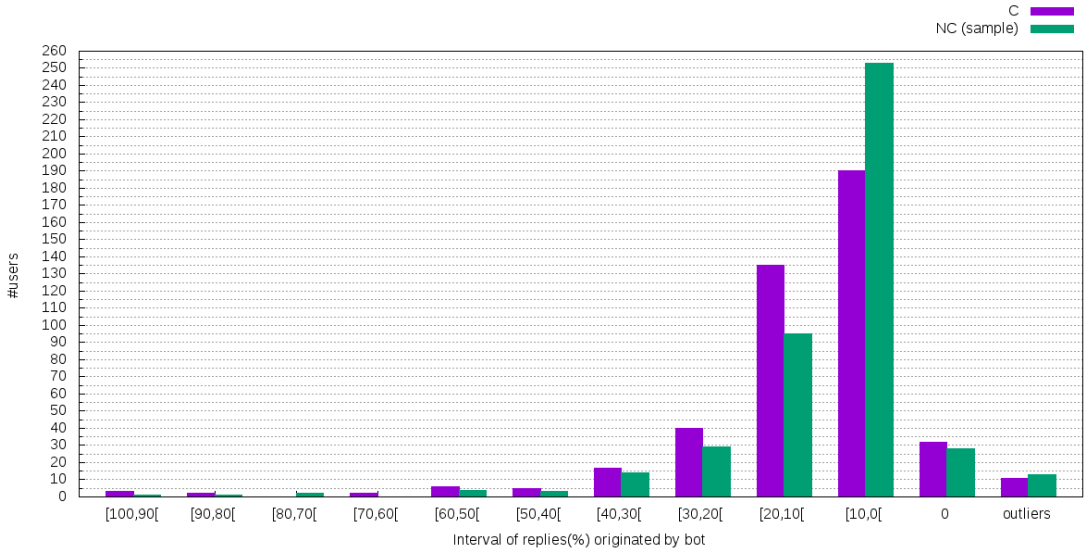


(A) Percentage of replies to bot's tweets posted by C and NC (sample) users.

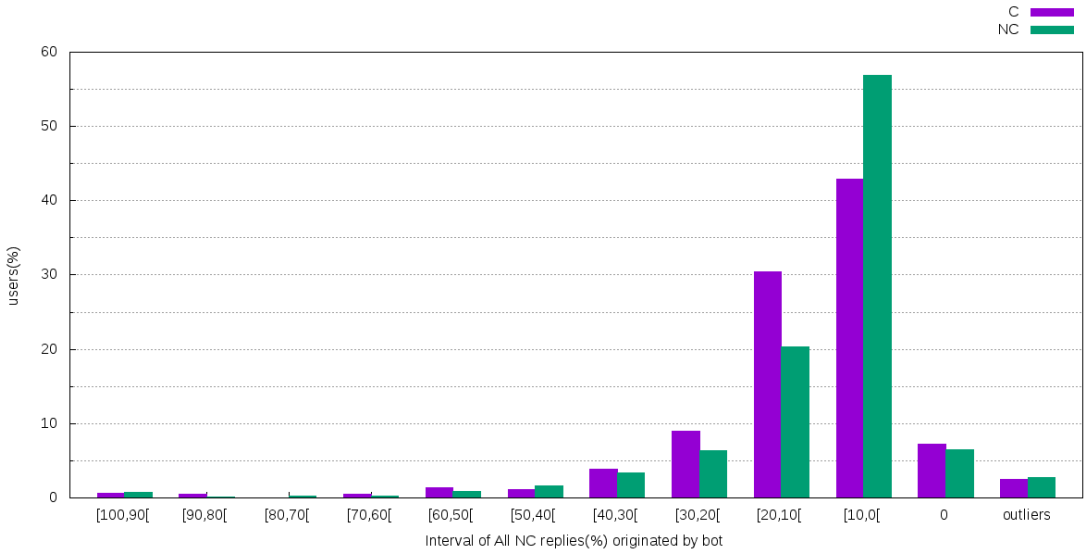


(B) % of populations w.r.t. the % of replies to bot's tweets.

FIGURE B.3: Comparative analysis between C and NC users w.r.t. the replies to bots' tweets. Here, the set of C users includes 443 humans (namely, *cut946*).



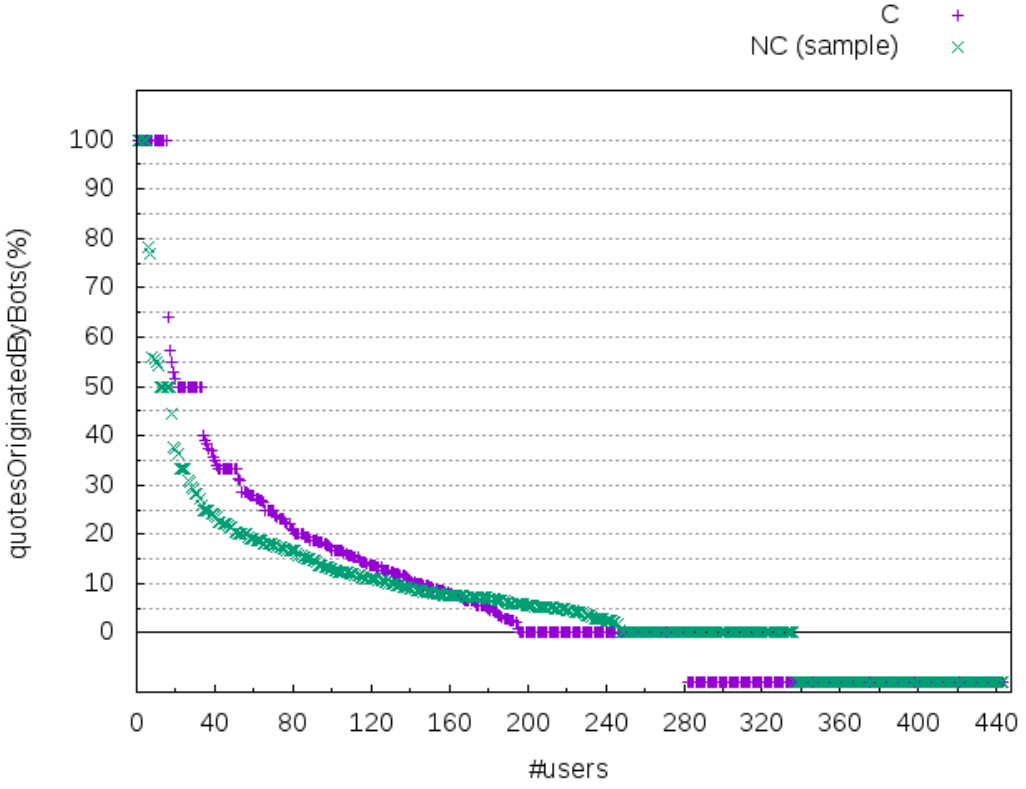
(A) Deciles of Figure B.3a.



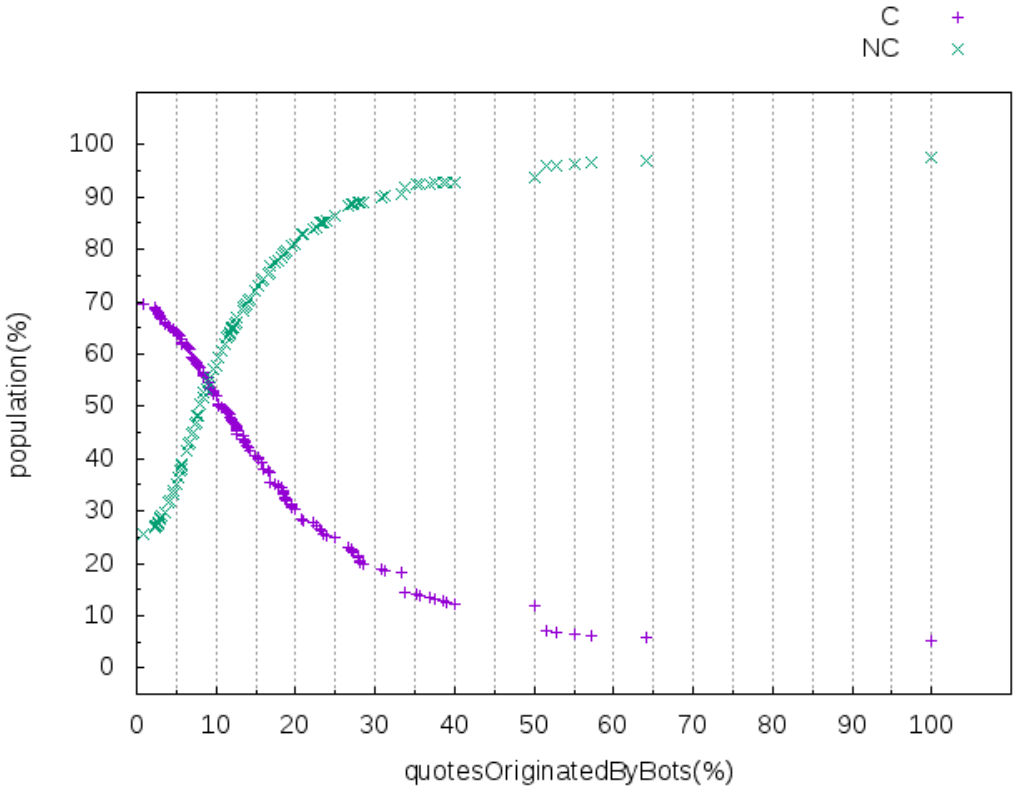
(B) Deciles of C and all NC users.

FIGURE B.4: Analysis using deciles – C *vs.* NC users w.r.t. the replies to bots' tweets. Here,the set of C users includes 443 humans (namely, *cut946*).

B.1.3 Quotes

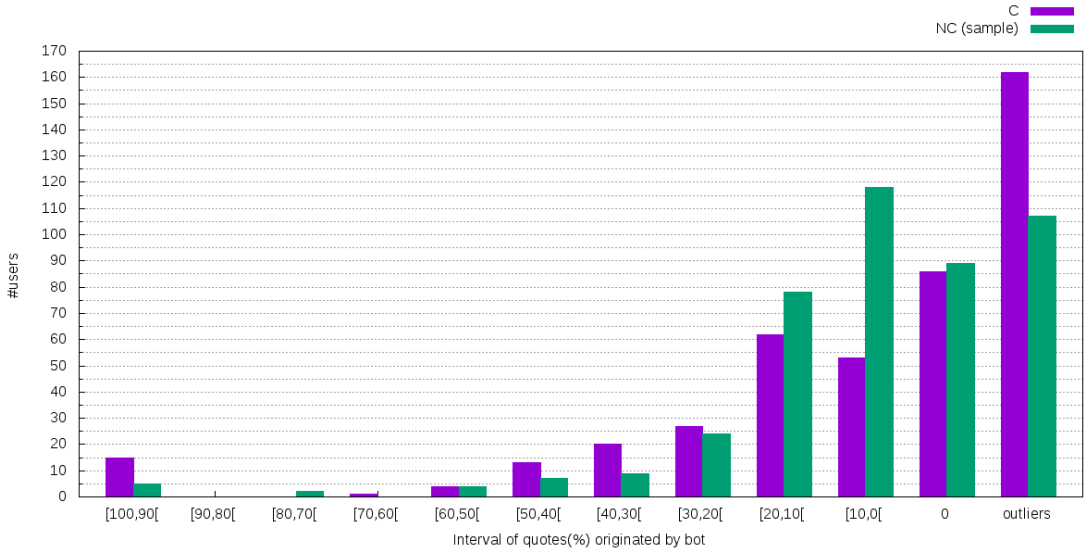


(A) Percentage of 'byBots'-quotes posted by C and NC (sample) users.

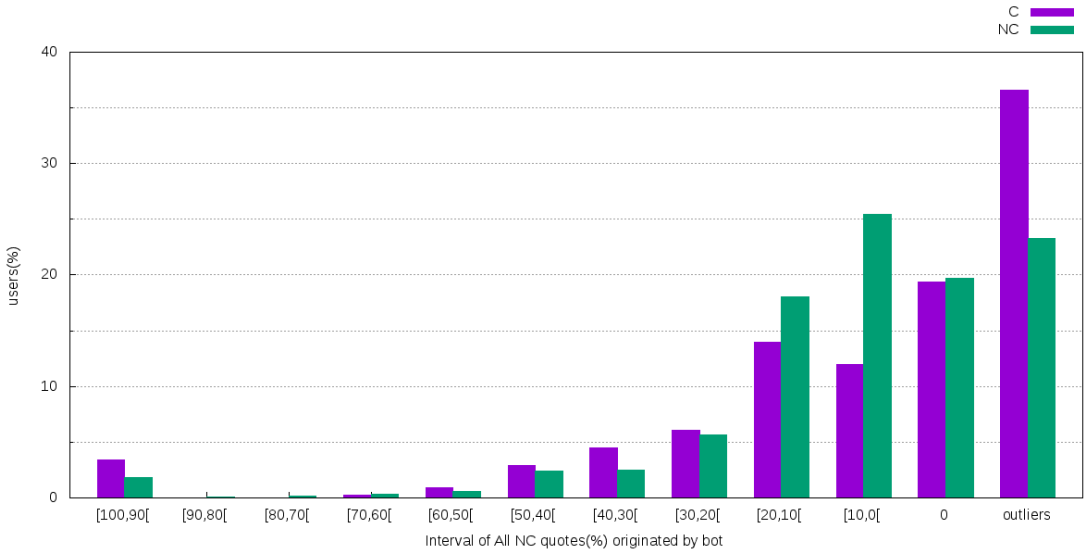


(B) % of populations w.r.t. the % of 'byBots'-quotes.

FIGURE B.5: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes. Here, the set of C users includes 443 humans (namely, *cut946*).



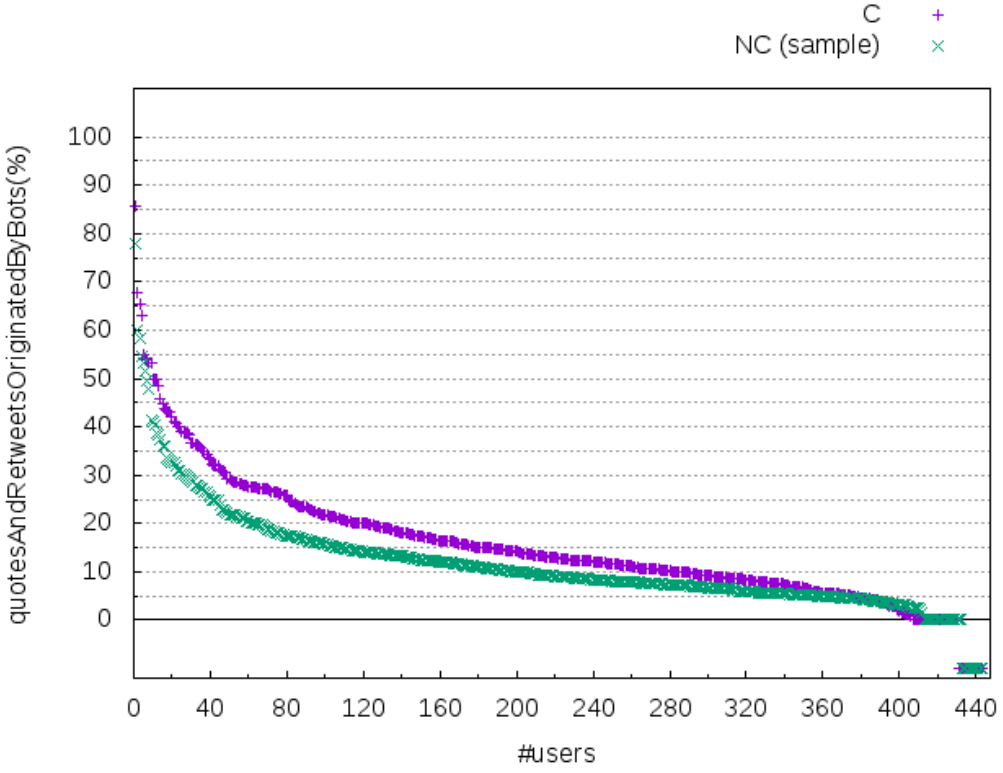
(A) Deciles of Figure B.5a



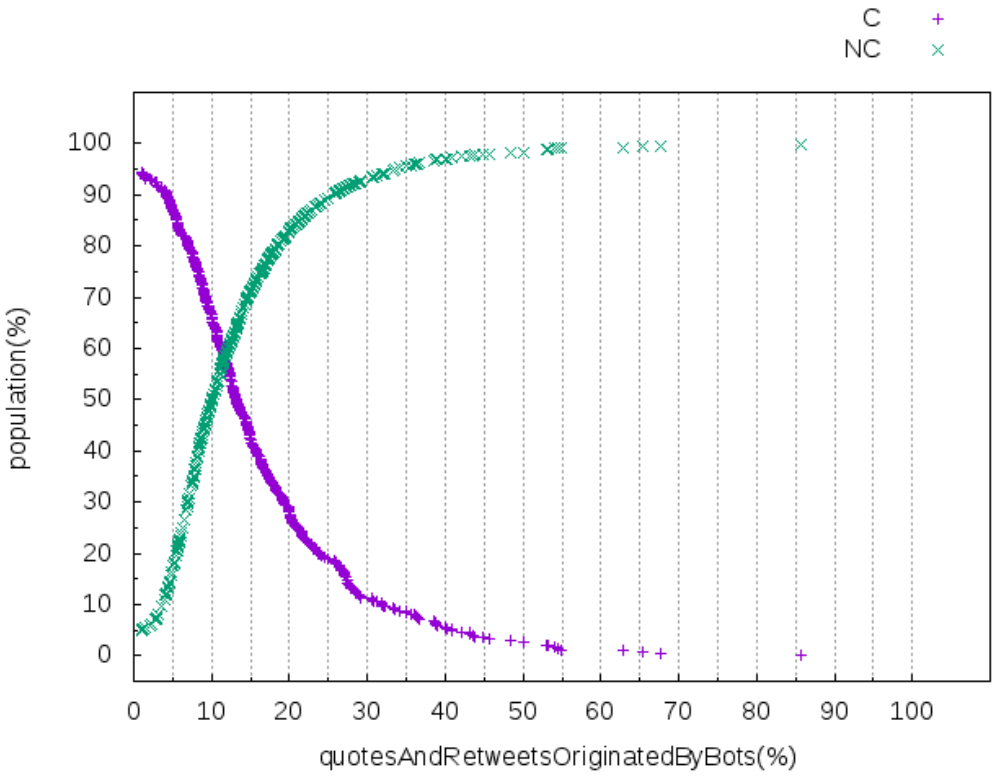
(B) Deciles of C and all NC users.

FIGURE B.6: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes. Here, the set of C users includes 443 humans (namely, *cut946*).

B.1.4 Retweets and quotes: aggregation

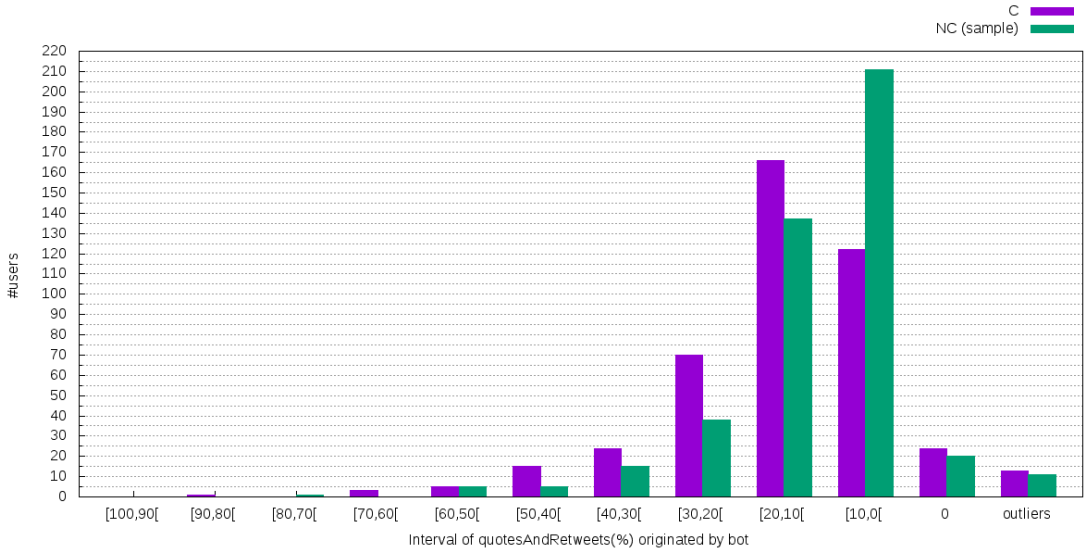


(A) Percentage of 'byBots'-quotes and retweets (jointly) posted by C and NC (sample) users.

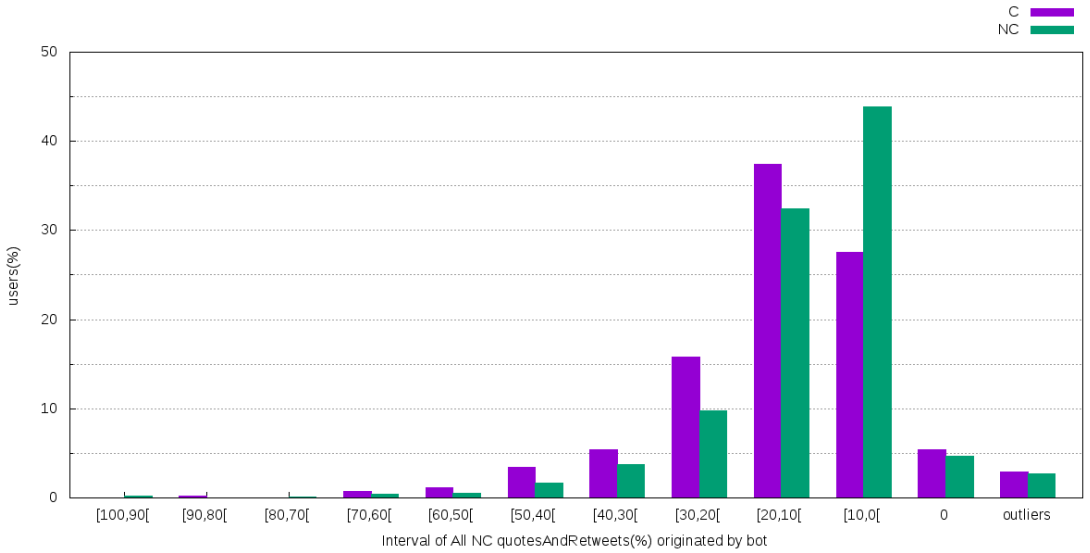


(B) % of populations w.r.t. the % of 'byBots'-quotes and retweets (jointly).

FIGURE B.7: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes and retweets (jointly). Here, the set of C users includes 443 humans (namely, *cut946*).



(A) Deciles of Figure B.7a

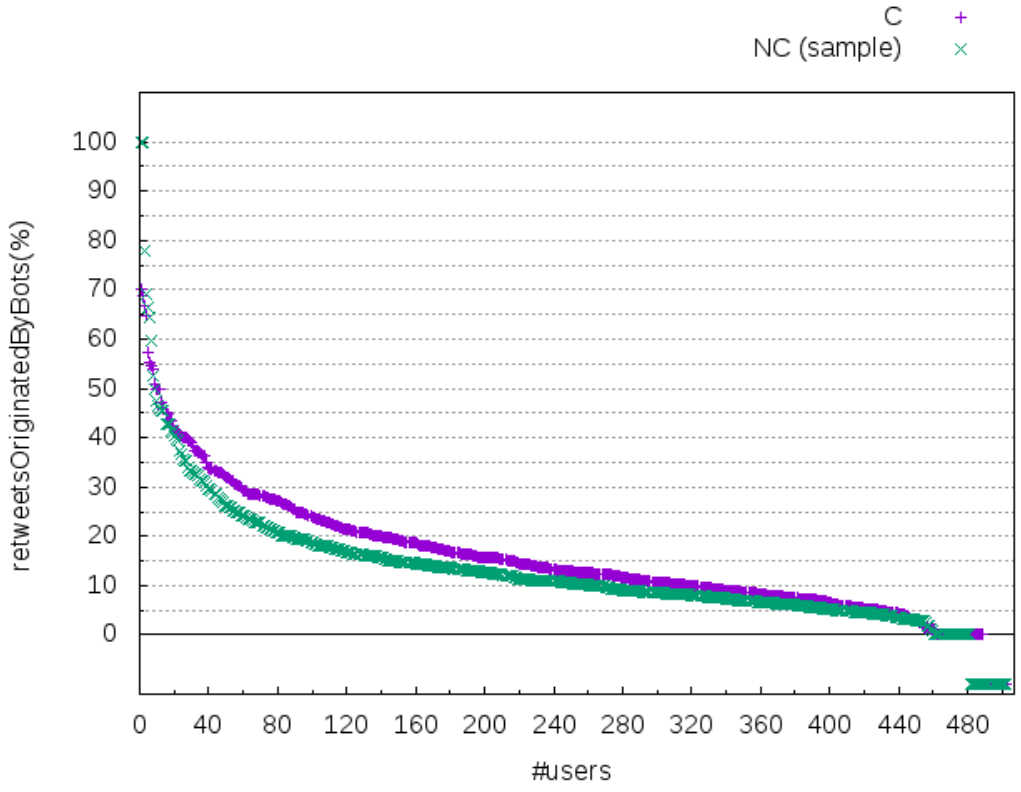


(B) Deciles of C and all NC users.

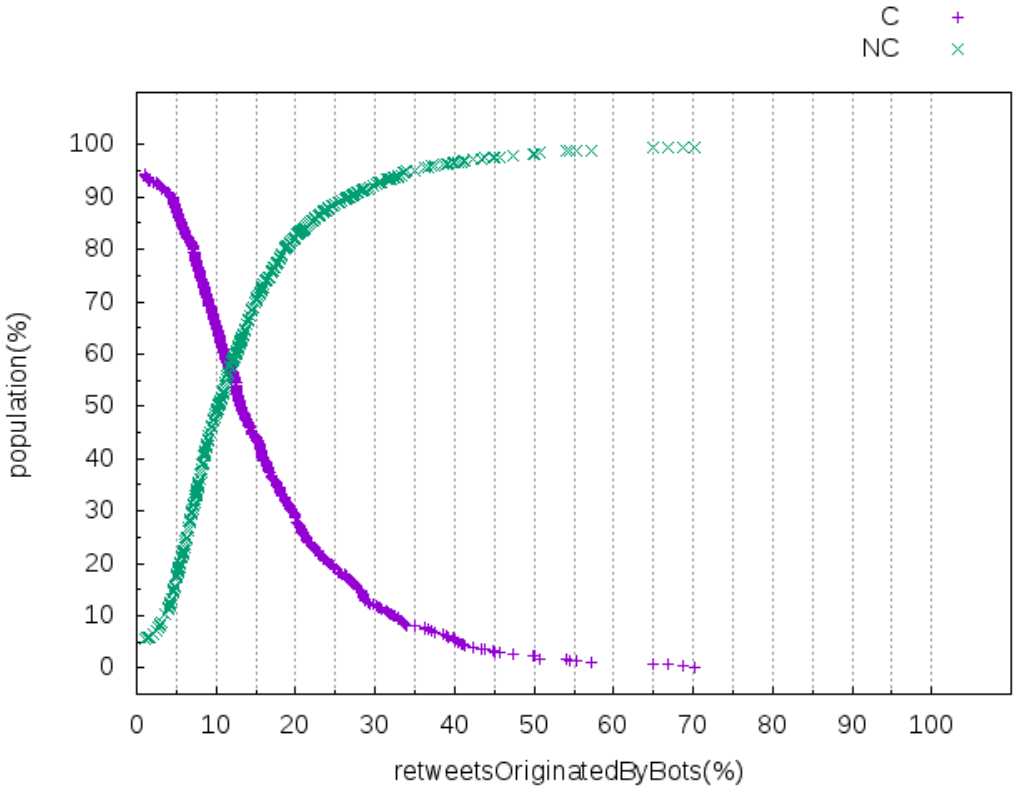
FIGURE B.8: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes and retweets (jointly). Here, the set of C users includes 443 humans (namely, *cut946*).

B.2 *cut1030*

B.2.1 Retweets

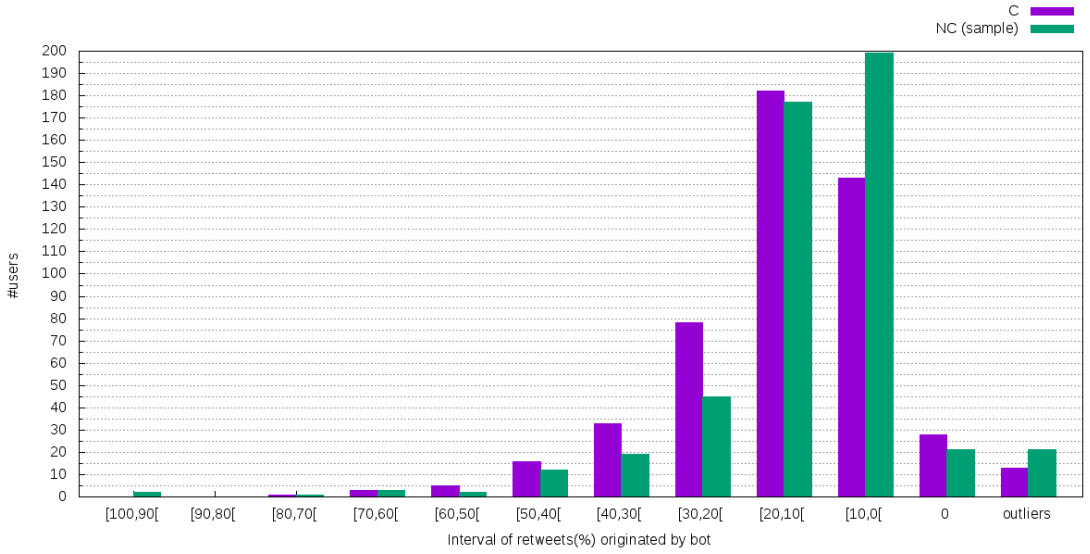


(A) Percentage of 'byBots'-retweets posted by C and NC (sample) users.

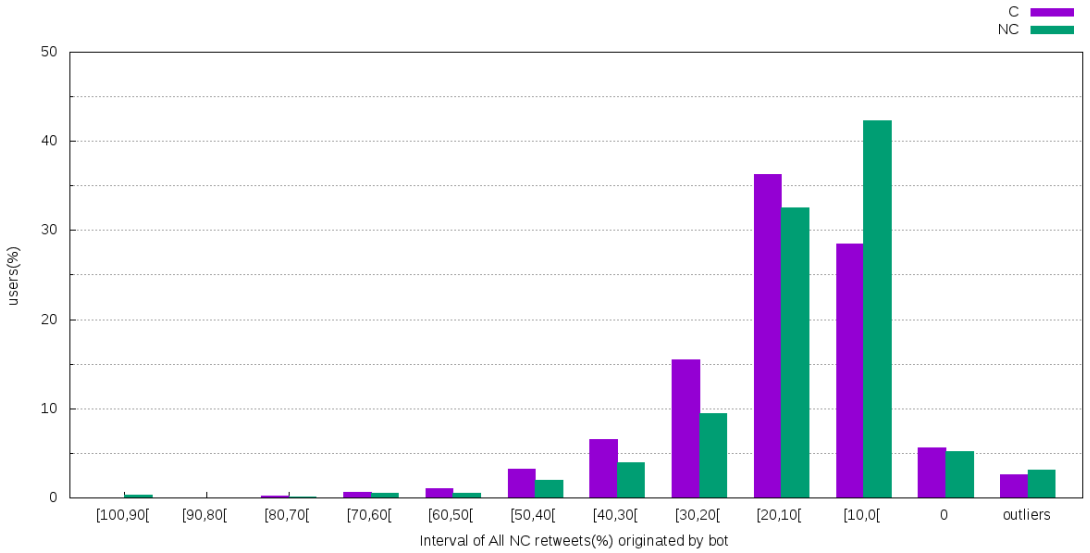


(B) % of populations w.r.t. the % of 'byBots'-retweets.

FIGURE B.9: Comparative analysis between C and NC users w.r.t. 'byBots'-retweets. Here, the set of C users includes 502 humans (namely, *cut1030*).



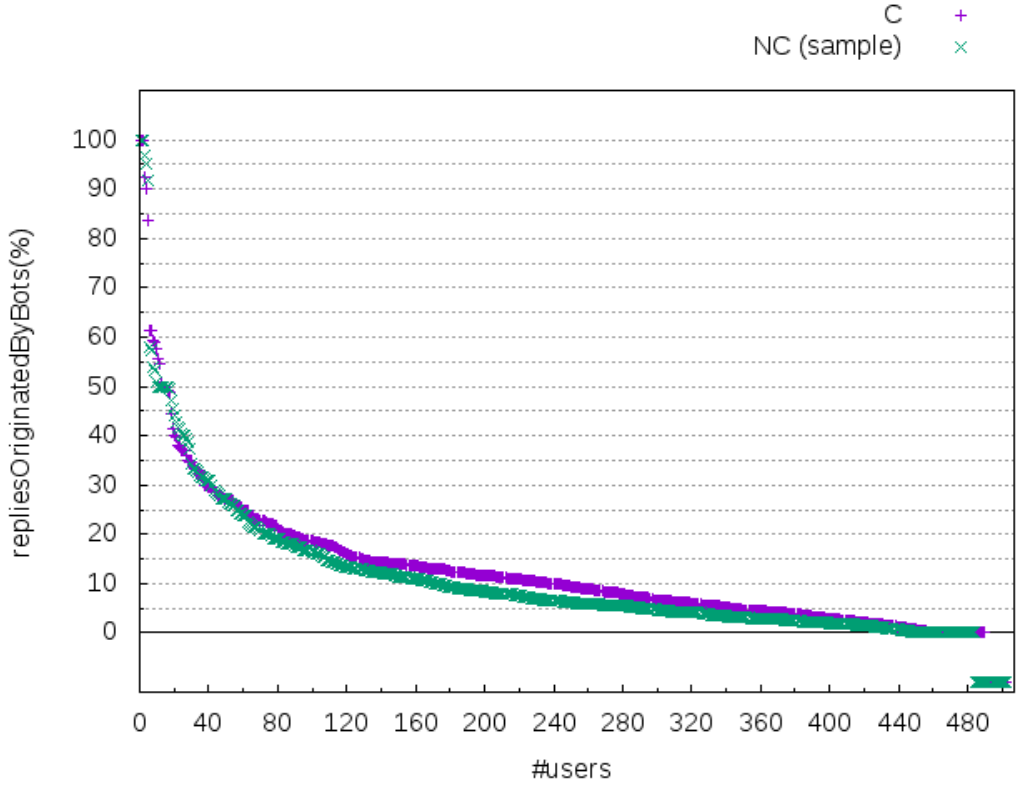
(A) Deciles of Figure B.9a



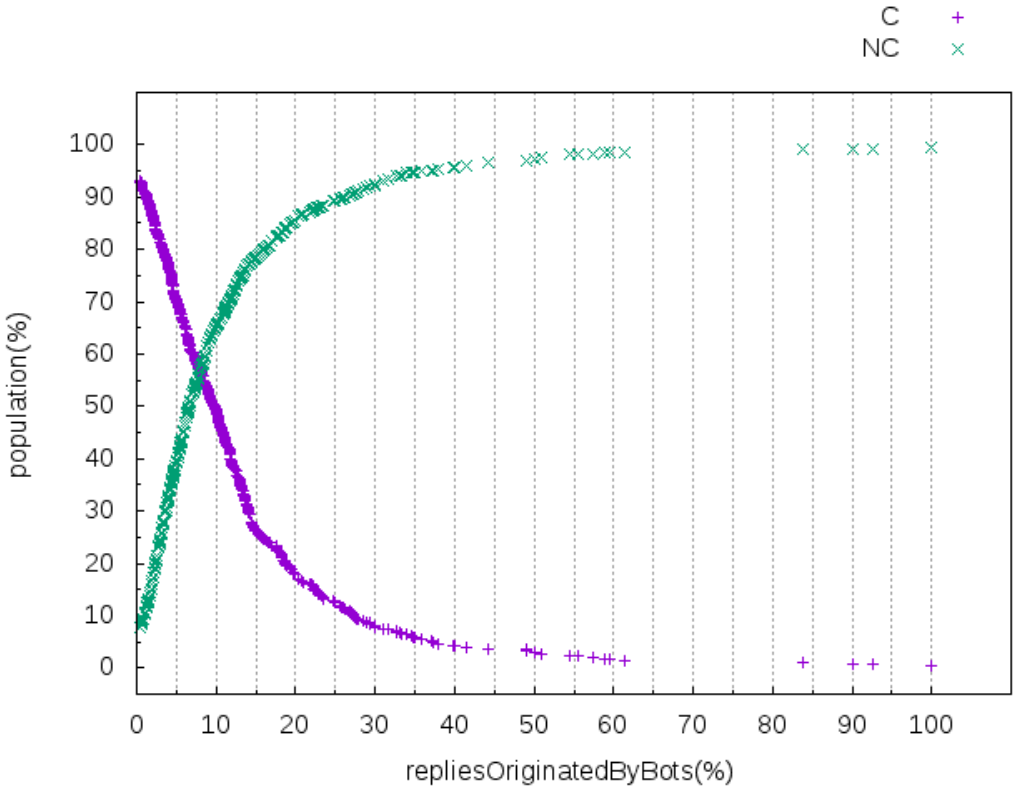
(B) Deciles of C and all NC users.

FIGURE B.10: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-retweets. Here, the set of C users includes 502 humans (namely, *cut1030*).

B.2.2 Replies

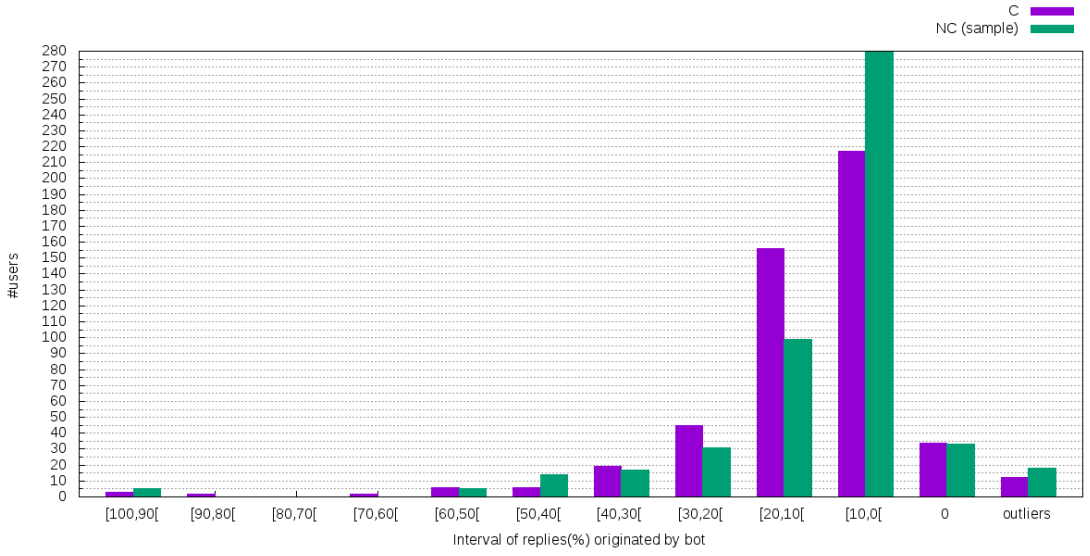


(A) Percentage of replies to tweets originated by bots.

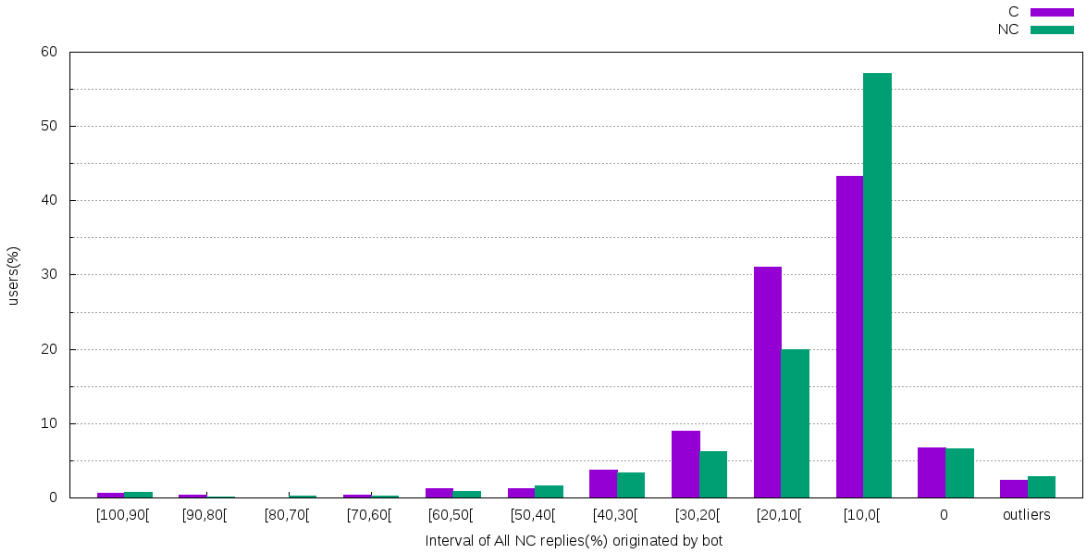


(B) % of populations w.r.t. the % of replies to bot's tweets.

FIGURE B.11: Comparative analysis between C and NC users w.r.t. the replies to bots' tweets. Here, the set of C users includes 502 humans (namely, *cut1030*).



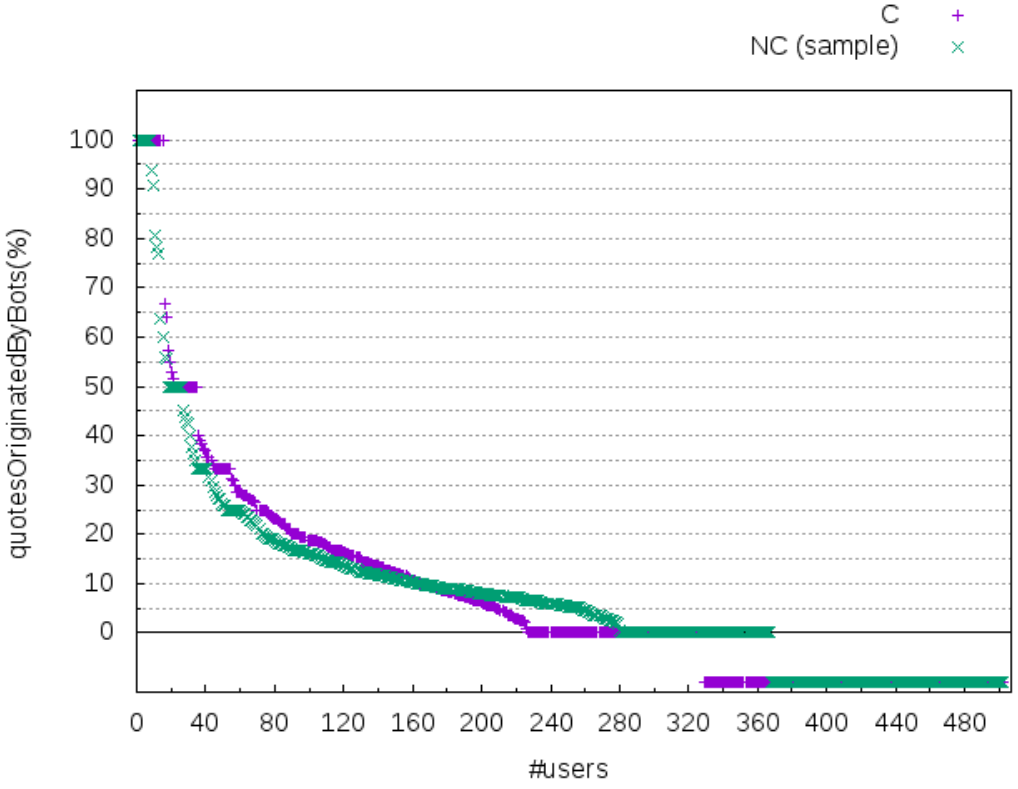
(A) Deciles of Figure B.11a.



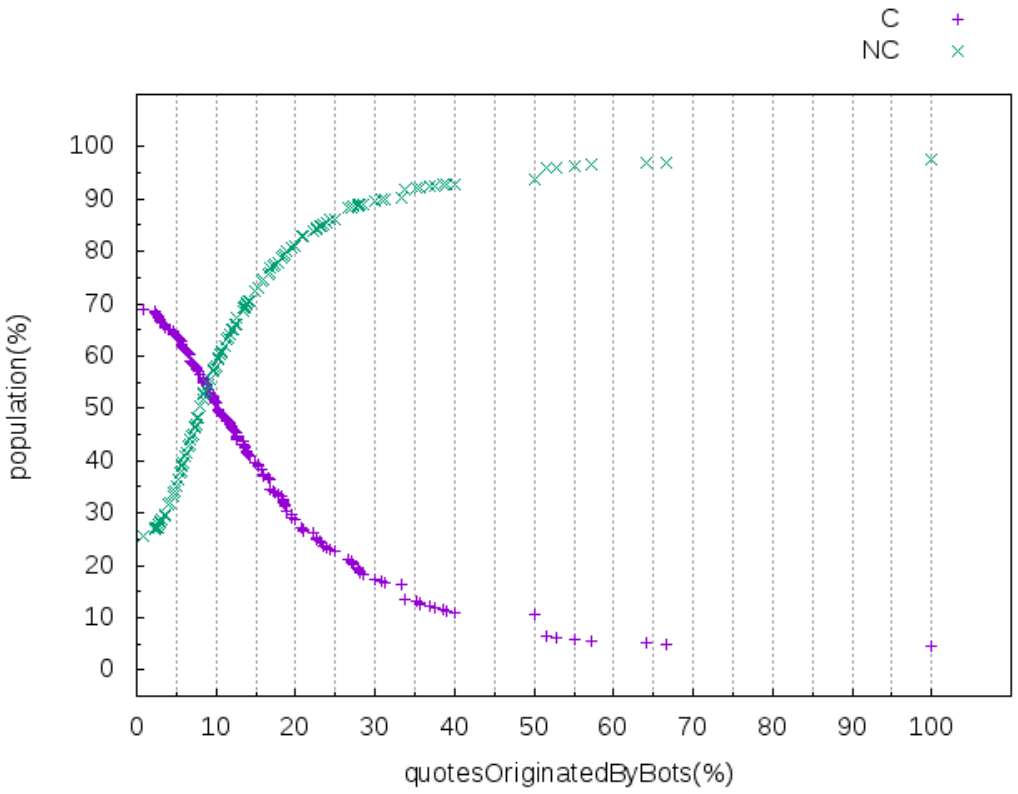
(B) Deciles of C and all NC users.

FIGURE B.12: Analysis using deciles – C *vs.* NC users w.r.t. the replies to bots' tweets. Here,the set of C users includes 502 humans (namely, *cut1030*).

B.2.3 Quotes

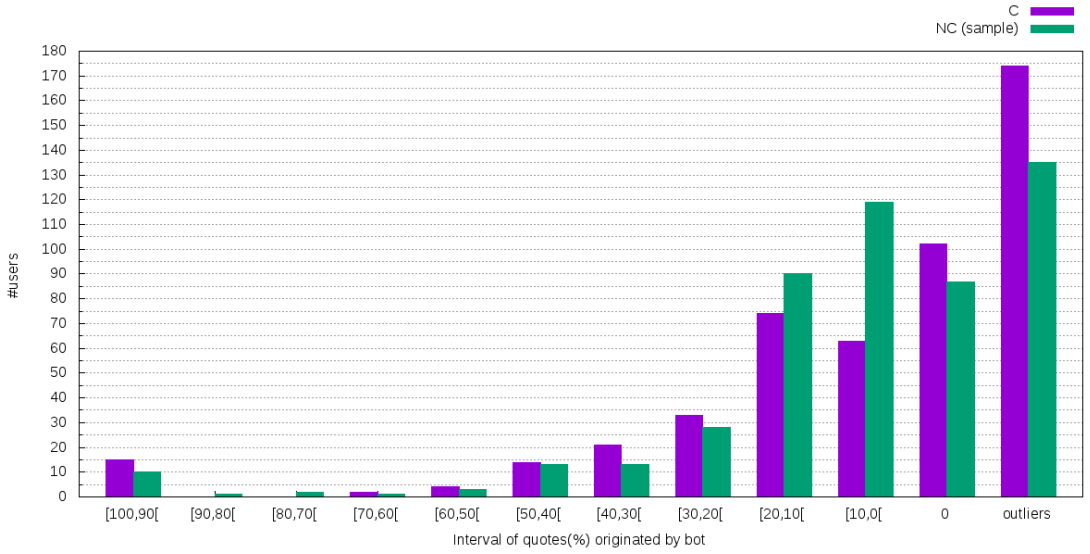


(A) Percentage of 'byBots'-quotes posted by C and NC (sample) users.

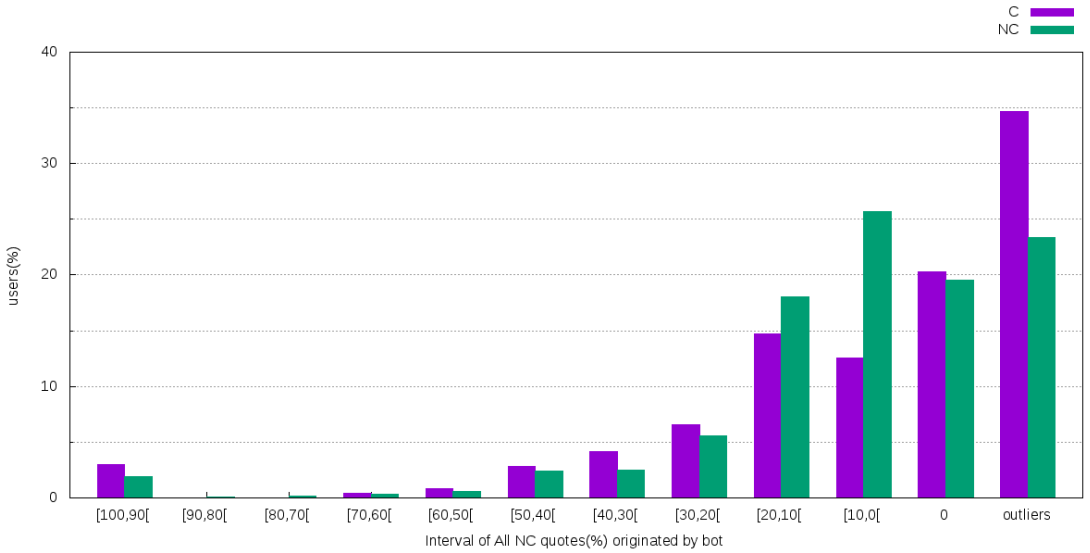


(B) % of populations w.r.t. the % of 'byBots'-quotes.

FIGURE B.13: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes. Here, the set of C users includes 502 humans (namely, *cut1030*).



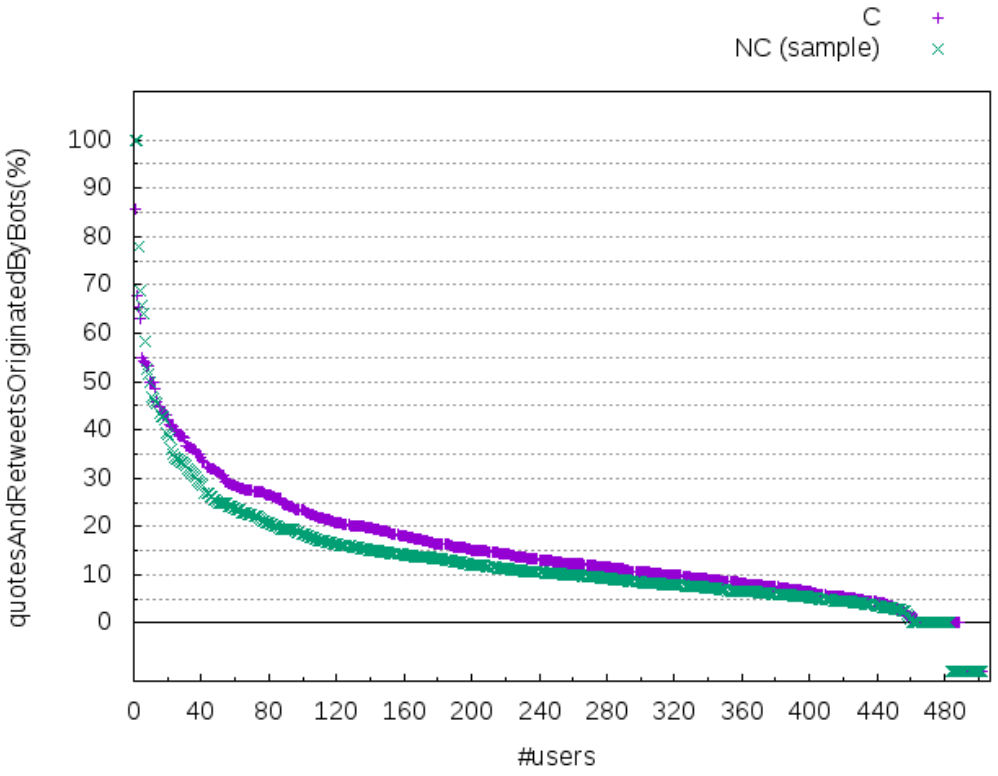
(A) Deciles of Figure B.13a



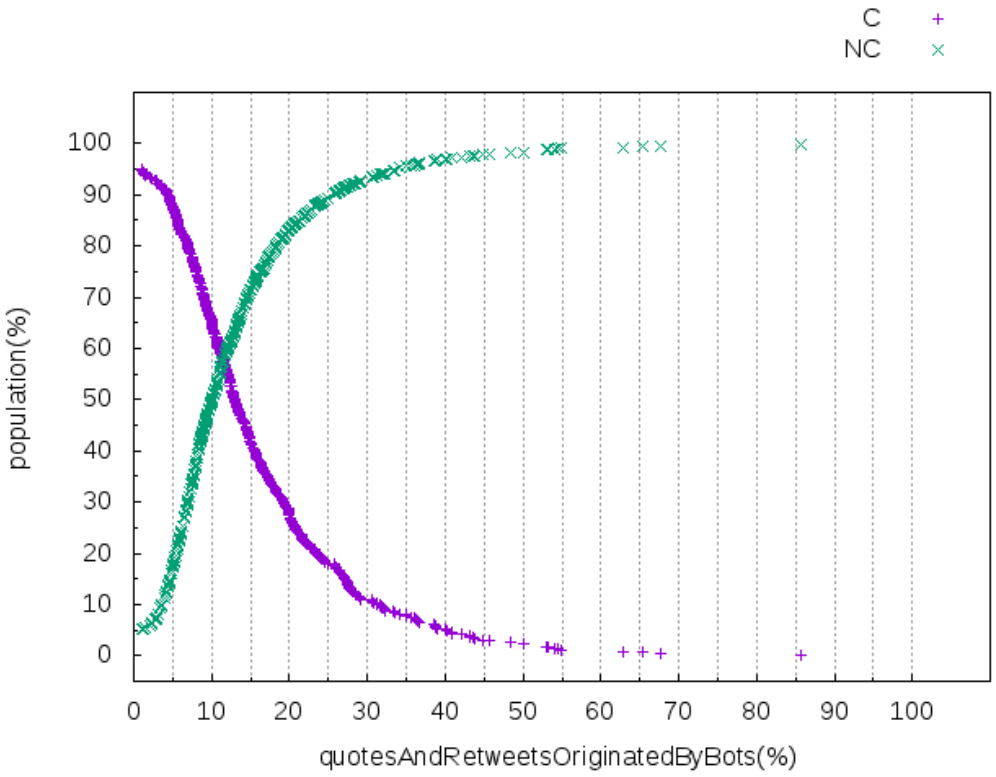
(B) Deciles of C and all NC users.

FIGURE B.14: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes. Here, the set of C users includes 502 humans (namely, *cut1030*).

B.2.4 Retweets and quotes: aggregation

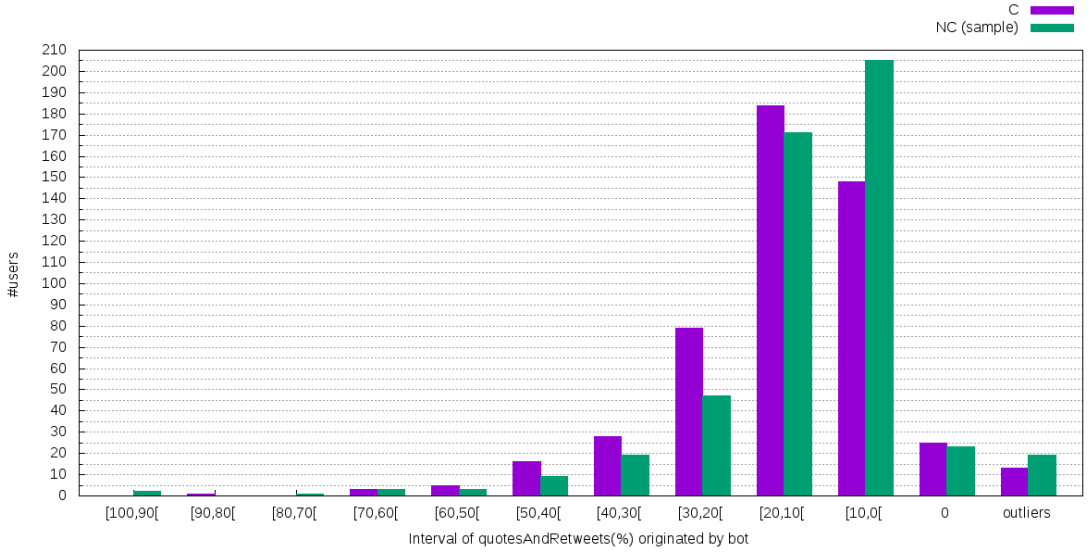


(A) Percentage of 'byBots'-quotes and retweets (jointly) posted by C and NC (sample) users.

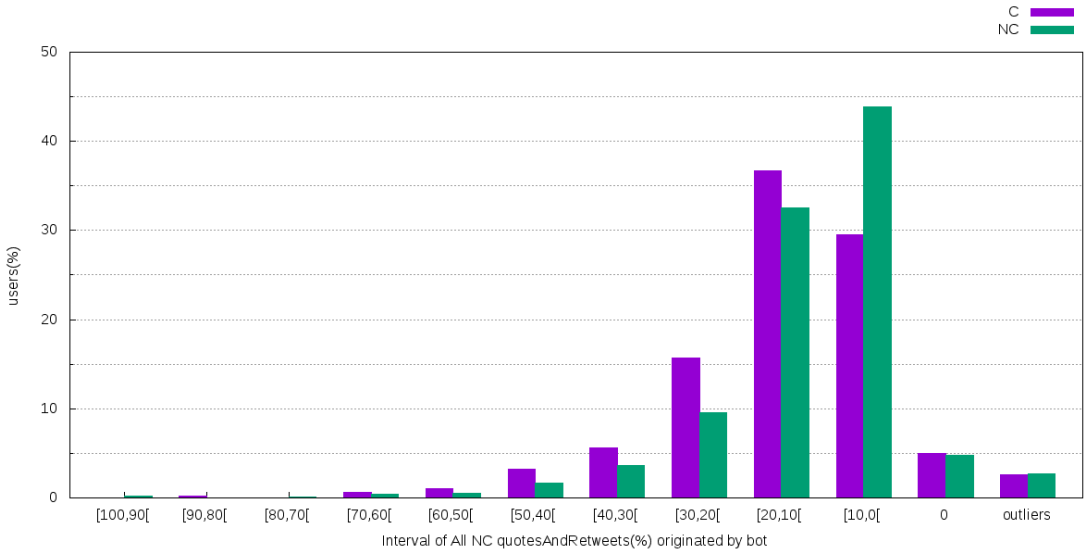


(B) % of populations w.r.t. the % of 'byBots'-quotes and retweets (jointly).

FIGURE B.15: Comparative analysis between C and NC users w.r.t. 'byBots'-quotes and retweets (jointly). Here, the set of C users includes 502 humans (namely, *cut1030*).



(A) Deciles of Figure B.15a



(B) Deciles of C and all NC users.

FIGURE B.16: Analysis using deciles – C *vs.* NC users w.r.t. 'byBots'-quotes and retweets (jointly). Here, the set of C users includes 502 humans (namely, *cut1030*).

Bibliography

[1] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic T. Stahl. A survey of data mining techniques for social media analysis. *J. Data Min. Digit. Humanit.*, 2014, 2014.

[2] David W. Aha, Dennis F. Kibler, and Marc K. Albert. Instance-based learning algorithms. *Mach. Learn.*, 6:37–66, 1991.

[3] Thomas Aichner and Frank Jacob. Measuring the degree of corporate social media use. *Int. J. Mark. Res.*, 57(2):257–276, 2015.

[4] Shehu Amina, Raúl Vera, Tooska Dargahi, and Ali Dehghantanha. A bibliometric analysis of botnet detection techniques. In *Handbook of Big Data and IoT Security*, pages 345–365. Springer, 2019.

[5] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning for control. *Artif. Intell. Rev.*, 11(1-5):75–113, 1997.

[6] Marco Avvenuti, Salvatore Bellomo, Stefano Cresci, Mariantonietta Noemi La Polla, and Maurizio Tesconi. Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In *WWW (Companion Volume)*, pages 1413–1421. ACM, 2017.

[7] Henning Baars and Hans-Georg Kemper. Management support with structured and unstructured data - an integrated business intelligence framework. *IS Management*, 25(2):132–148, 2008.

[8] Alessandro Balestrucci. How many bots are you following? In *ITASEC*, volume 2597 of *CEUR Workshop Proceedings*, pages 47–59. CEUR-WS.org, 2020.

[9] Alessandro Balestrucci, Rocco De Nicola, Petrocchi Marinella, and Trubiani Catia. Emergent properties of bots-humans relationships through behavioral analysis of credulous twitter users. *Inf. Process. Manag.*, 2020 (under review).

[10] Alessandro Balestrucci and Rocco De Nicola. Credulous users and fake news: a real case study on the propagation in twitter. In *EAIS*, pages 1–8. IEEE, 2020.

[11] Alessandro Balestrucci, Rocco De Nicola, Omar Inverso, and Catia Trubiani. Identification of credulous users on twitter. In *SAC*, pages 2096–2103. ACM, 2019.

[12] Alessandro Balestrucci, Rocco De Nicola, Marinella Petrocchi, and Catia Trubiani. Do you really follow them? automatic detection of credulous twitter users. In *IDEAL (1)*, volume 11871 of *Lecture Notes in Computer Science*, pages 402–410. Springer, 2019.

[13] Albert A. Barreda, Khaldoon Nusair, Youcheng Wang, Fevzi Okumus, and Anil Bilgihan. The impact of social media activities on brand image and emotional attachment. *J. Hosp. Tour. Technol.*, 2020.

[14] Larry M. Bartels. Beyond the running tally: Partisan bias in political perceptions. *Polit. Behav.*, 24(2):117–150, 2002.

[15] Marco T. Bastos and Dan Mercea. The brexit botnet and user-generated hyperpartisan news. *Soc. Sci. Comput. Rev.*, 37(1):38–54, 2019.

[16] Christoph Besel, Jörg Schlötterer, and Michael Granitzer. Inferring semantic interest profiles from twitter followees: does twitter know better than your friends? In *SAC*, pages 1152–1157, 2016.

[17] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11-7), 2016.

[18] Sajid Yousuf Bhat and Muhammad Abulaish. Community-based features for identifying spammers in online social networks. In *ASONAM*, pages 100–107. ACM, 2013.

[19] Grant Blank and Bianca C Reisdorf. The participatory web: A user perspective on web 2.0. *Inf. Commun. Soc.*, 15(4):537–554, 2012.

[20] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Inf. Sci.*, 497:38–55, 2019.

[21] Samantha Bradshaw and Philip N. Howard. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. *Computational Propaganda Research Project – Oxford Internet Institute*, page 37, 2017.

[22] Samantha Bradshaw and Philip N Howard. The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation. *University of Oxford*, page 25, 2019.

[23] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[24] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[25] Scott J. Brennen, Felix M. Simon, Philip N. Howard, and Rasmus-Kleis Nielsen. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 2020.

[26] Zhan Bu, Zhengyou Xia, and Jiandong Wang. A sock puppet detection algorithm on virtual spaces. *Knowl. Based Syst.*, 37:366–377, 2013.

[27] Talha Burki. Vaccine misinformation and social media. *Lancet Digital Health*, 1(6):e258–e259, 2019.

[28] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Comput. Electr. Eng.*, 40(1):16–28, 2014.

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

[30] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[31] Jing Chen, Long Cheng, Xi Yang, Jun Liang, Bing Quan, and Shoushan Li. Joint learning with both classification and regression models for age prediction. In *J. Phys.: Conf. Ser. 1168 032016*, volume 1168, 2019.

[32] Man Lai Cheung, Guilherme D. Pires, and Philip J. Rosenberger III. Developing a conceptual model for examining social media marketing effects on brand awareness and brand image. *Int. J. Econ. Bus. Res.*, 17(3):243–261, 2019.

[33] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *ACSAC*, pages 21–30. ACM, 2010.

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

[35] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

[36] Aaron Clauset. A brief primer on probability distributions. In *Santa Fe Institute*, 2011.

[37] William W. Cohen. Fast effective rule induction. In *ICML*, pages 115–123. Morgan Kaufmann, 1995.

[38] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In *ASIST*, volume 52, pages 1–4. Wiley, 2015.

[39] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[40] Irene Costera Meijer and Tim Groot Kormelink. Checking, sharing, clicking and linking: Changing patterns of news use between 2004 and 2014. *Digit. Journal.*, 3(5):664–679, 2015.

[41] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. *J. Big Data*, 2:23, 2015.

[42] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *WWW (Companion Volume)*, pages 963–972. ACM, 2017.

[43] Stefano Cresci, Marinella Petrocchi, Angelo Spognardi, and Stefano Tognazzi. On the capability of evolved spambots to evade detection via genetic engineering. *Online Soc. Networks Media*, 9:1–16, 2019.

[44] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: Efficient detection of fake twitter followers. *Decis. Support Syst.*, 80:56–71, 2015.

[45] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intell. Syst.*, 31(5):58–64, 2016.

[46] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Social fingerprinting: Detection of spambot groups through dna-inspired behavioral modeling. *IEEE Trans. Dependable Secur. Comput.*, 15(4):561–576, 2018.

[47] Florian Daniel, Cinzia Cappiello, and Boualem Benatallah. Bots acting like humans: Understanding and preventing harm. *IEEE Internet Comput.*, 23(2):40–49, 2019.

[48] Isaac David, Oscar S. Siordia, and Daniela Moctezuma. Features combination for the detection of malicious twitter accounts. In *IEEE ROPEC*, pages 1–6, 2016.

[49] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *WWW (Companion Volume)*, pages 273–274. ACM, 2016.

[50] O.V. Deryugina. Chatterbots. *Sci. Tech. Inf. Process.*, 37(2):143–147, 2010.

[51] Phillip George Efthimion, Scott Payne, and Nicholas Proferes. Supervised machine learning bot detection techniques to identify social twitter bots. *SMU Data Science Review*, 1(2):5, 2018.

[52] Frank Eibe, M.A. Hall, and I.H. Witten. *The WEKA Workbench. Online Appendix for" Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 2016.

[53] Cerchia Alina Elena. Social media–a strategy in developing customer relationship management. *Procedia Econ.*, 39:785–790, 2016.

[54] Robert J. Elliott, Lakhdar Aggoun, and John B. Moore. *Hidden Markov models: estimation and control*, volume 29. Springer Science & Business Media, 2008.

[55] Robert M Entman, Jörg Matthes, and Lynn Pellicano. Nature, sources, and effects of news framing. *The handbook of journalism studies*, pages 175–190, 2009.

[56] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

[57] Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Cinzia Squicciarini. Uncovering crowdsourced manipulation of online reviews. In *SIGIR*, pages 233–242. ACM, 2015.

[58] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34, 1996.

[59] Miriam Fernández and Harith Alani. Online misinformation: Challenges and future directions. In *WWW (Companion Volume)*, pages 595–602. ACM, 2018.

[60] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8), 2017.

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

[62] Eibe Frank and Remco R. Bouckaert. Conditional density estimation with class probability estimators. In *ACML*, volume 5828 of *Lecture Notes in Computer Science*, pages 65–81. Springer, 2009.

[63] Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3):277–296, 1999.

[64] Jerome H Friedman. Stochastic gradient boosting. *Comput. Stat. Data An.*, 38(4):367–378, 2002.

[65] Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163, 1997.

[66] Christina Gangware and William Nemr. *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*. Park Advisors, 2019.

[67] Maria Giatsoglou, Despoina Chatzakou, Neil Shah, Alex Beutel, Christos Faloutsos, and Athena Vakali. Nd-sync: Detecting synchronized fraud activities. In *PAKDD (2)*, volume 9078 of *Lecture Notes in Computer Science*, pages 201–214. Springer, 2015.

[68] Zafar Gilani, Reza Farahbakhsh, and Jon Crowcroft. Do bots impact twitter activity? In *WWW (Companion Volume)*, pages 781–782. ACM, 2017.

[69] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, and Jon Crowcroft. A large-scale behavioural analysis of bots and humans on twitter. *ACM Trans. Web*, 13(1):7:1–7:23, 2019.

[70] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.

[71] Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. *Policy Internet*, 2018.

[72] Mark. S. Granovetter. The strength of weak ties. *AJS*, 78(6):1360–1380, 1973.

[73] Imene Guellil and Kamel Boukhalfa. Social big data mining: A survey focused on opinion mining and sentiments analysis. In *ISPS*, pages 1–10. IEEE, 2015.

[74] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[75] Stefanie Haustein, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *J. Assoc. Inf. Sci. Technol.*, 67(1), 2016.

[76] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, pages 1322–1328. IEEE, 2008.

[77] Tin Kam Ho. Random decision forests. In *ICDAR*, pages 278–282. IEEE Computer Society, 1995.

[78] Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, Eibe Frank, and Mark A. Hall. Multiclass alternating decision trees. In *ECML*, volume 2430 of *Lecture Notes in Computer Science*, pages 161–172. Springer, 2002.

[79] Robert C. Holte. Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.*, 11:63–91, 1993.

[80] Philip N. Howard and Bence Kollanyi. Bots, #strongerin, and #brexit: Computational propaganda during the UK-EU referendum. *Available at SSRN 2798311*, abs/1606.06356, 2016.

[81] David C. Howell. *Statistical methods for psychology*. Cengage Learning, 2009.

[82] Henry Hsu and Peter A. Lachenbruch. Paired t test. *Encyclopedia of Biostatistics*, 6, 2005.

[83] Geoff Hulten, Laurie Spencer, and Pedro M. Domingos. Mining time-changing data streams. In *KDD*, pages 97–106. ACM, 2001.

[84] Wayne Iba and Pat Langley. Induction of one-level decision trees. In *ML*, pages 233–240. Morgan Kaufmann, 1992.

[85] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *UAI*, pages 338–345. Morgan Kaufmann, 1995.

[86] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Bus. Horiz.*, 53(1):59–68, 2010.

[87] Anna Kata. A postmodern pandora's box: anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716, 2010.

[88] Tobias R. Keller and Ulrike Klinger. Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Polit. Commun.*, 36(1):171–189, 2019.

[89] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Theory Comput.*, 11:105–147, 2015.

[90] John T. Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[91] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[92] Kenneth A. Lachlan, Zhan Xu, Emily E. Hutter, Rainear Adam, and Patric R. Spence. A little goes a long way: serial transmission of twitter content associated with hurricane irma and implications for crisis communication. *JSIS*, 14(1):16–26, 2019.

[93] Niels Landwehr, Mark A. Hall, and Eibe Frank. Logistic model trees. *Mach. Learn.*, 59(1-2):161–205, 2005.

[94] David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[95] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *J. R. Stat. Soc. C-Appl.*, 41(1):191–201, 1992.

[96] Jong Bum Lee and Jee-Hyong Lee. An iterative undersampling of extremely imbalanced data using CSVM. In *ICMV*, volume 9445 of *SPIE Proceedings*, page 94452B. SPIE, 2014.

[97] Kyumin Lee, Prithivi Tamilarasan, and James Caverlee. Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media. In *ICWSM*. The AAAI Press, 2013.

[98] Kyumin Lee, Steve Webb, and Hancheng Ge. Characterizing and automatically detecting crowdturfing in fiverr and twitter. *Social Netw. Analys. Mining*, 5(1):2:1–2:16, 2015.

[99] Martin C. Libicki. What is information warfare? Technical report, National Defense University – Institute for National Strategic Studies, 1995.

[100] Martin C. Libicki. *Information dominance*. National Defense University, Institute for National Strategic Studies, 1997.

[101] Hubert W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.*, 62(318):399–402, 1967.

[102] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Inf. Sci.*, 409:17–26, 2017.

[103] Shenghua Liu, Bryan Hooi, and Christos Faloutsos. Holoscope: Topology-and-spike aware fraud detection. In *CIKM*, pages 1539–1548. ACM, 2017.

[104] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B*, 39(2):539–550, 2009.

[105] Tetyana Lokot and Nicholas Diakopoulos. News Bots: Automating news and information dissemination on Twitter. *Digit. Journal.*, 4(6), 2016.

[106] Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. Red bots do it better: Comparative analysis of social bot partisan behavior. In *WWW (Companion Volume)*, pages 1007–1012. ACM, 2019.

[107] Theo Lynn, Philip D. Healy, Steven Kilroy, Graham Hunt, Lisa van der Werff, Shankar Venkatagiri, and John P. Morrison. Towards a general research framework for social media research using big data. In *IPCC*, pages 1–8. IEEE, 2015.

[108] David J.C. MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

[109] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[110] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[111] Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta - Protein Structure*, 405(2):442–451, 1975.

[112] Jacinta Mwende Maweu. "Fake Elections"? cyber propaganda, disinformation and the 2017 general elections in kenya. *African Journal. Stud.*, pages 1–15, 2020.

[113] Aaron M. McCright and Riley E. Dunlap. The politicization of climate change and polarization in the american public's views of global warming, 2001–2010. *Sociol. Q.*, 52(2):155–194, 2011.

[114] Amanda J. Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. Botwalk: Efficient adaptive exploration of twitter bot networks. In *ASONAM*, pages 467–474. ACM, 2017.

[115] Tom M. Mitchell. *Machine learning, International Edition.* McGraw-Hill Series in Computer Science. McGraw-Hill, 1997.

[116] Silvia Mitter, Claudia Wagner, and Markus Strohmaier. A categorization scheme for socialbot attacks in online social networks. *arXiv preprint arXiv:1402.6288*, 2014.

[117] Jojo Moolayil and Suresh John. *Learn Keras for Deep Neural Networks.* Springer, 2019.

[118] Atif Mushtaq, Todd Rosenberry, Ashar Aziz, and Ali Islam. Distributed systems and methods for automatically detecting unknown bots and botnets, February 5 2019. US Patent 10,200,384.

[119] National Intelligence Council. Assessing Russian Activities and Intentions in Recent US Elections. *Intelligence Community Assessment*, 1(January):14, 2017.

[120] German Neubaum and Nicole C. Krämer. Opinion climates in social media: Blending mass and interpersonal communication. *Hum. Commun. Res.*, 43(4):464–476, 2017.

[121] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David Levy, and Rasmus Kleis Nielsen. Digital news report. *RISJ*, 2017.

[122] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David A.L. Levy, and Rasmus-Kleis Nielsen. Digital news report. *RISJ*, 2016.

[123] Nic Newman, Richard Fletcher, Anne Schulz, and Simge an Andı. Digital news report. *RISJ*, 2020.

[124] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David A.L. Levy, and Rasmus-Kleis Nielsen. Digital news report. *RISJ*, 2018.

[125] Jonathan A. Obar and Steve Wildman. Social media definition and the governance challenge: An introduction to the special issue. *Telecomm Policy*, 39(9):745–750, 2015.

[126] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. In *LREC*, pages 6086–6093, 2020.

[127] Sankar K. Pal and Sushmita Mitra. Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Networks*, 3(5):683–697, 1992.

[128] Nidhi A. Patel and Rakesh Patel. A survey on fake review detection using machine learning techniques. In *IEEE ICCCA*, pages 1–6, 2018.

[129] Fay Cobb Payton and Cherie Conley. Fear or danger threat messaging: The dark side of social media. In *Social Media: Global Perspectives, Applications and Benefits and Dangers*, pages 23–37. NOVA Publishers, 2014.

[130] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David Rand. Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv.com/3n9u8*, 2019.

[131] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, and David Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *PsyArXiv Preprints*, 10, 2020.

[132] Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.*, 116(7):2521–2526, 2019.

[133] Jesús M. Pérez, Javier Muguerza, Olatz Arbelaitz, Ibai Gurrutxaga, and José Ignacio Martín. Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognit. Lett.*, 28(4):414–422, 2007.

[134] Sarah Phillips. A brief history of facebook. *The Guardian*, 25(7):553–592, 2007.

[135] Guangyuan Piao and John G. Breslin. Inferring user interests for passive users on twitter by leveraging followee biographies. In *ECIR*, volume 10193 of *LNCS*, pages 122–133, 2017.

[136] J Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods — Support Vector Learning*, pages 185–208, 1999.

[137] J. Ross Quinlan. Simplifying decision trees. *Int. J. Man Mach. Stud.*, 27(3):221–234, 1987.

[138] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[139] Ross J. Quinlan. Learning with continuous classes. In *AusAI*, pages 343–348. World Scientific, 1992.

[140] Sam Rowlands. Misinformation on abortion. *Eur. J. Contracept. Reprod. Health Care*, 16(4):233–240, 2011.

[141] Stuart J. Russell and Peter Norvig. *Artificial intelligence - a modern approach, 2nd Edition*. Prentice Hall, 2003.

[142] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229, 1959.

[143] Prudhvi Ratna Badri Satya, Kyumin Lee, Dongwon Lee, Thanh Tran, and Jason (Jiasheng) Zhang. Uncovering fake likers in online social networks. In *CIKM*, pages 2365–2370. ACM, 2016.

[144] Saiph Savage, Andrés Monroy-Hernández, and Tobias Höllerer. Botivist: Calling volunteers to action using online bots. In *CSCW*, pages 811–820. ACM, 2016.

[145] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. *arXiv preprint arXiv:2006.06867*, 2020.

[146] James Schnebly and Shamik Sengupta. Random forest twitter bot classifier. In *CCWC*, pages 506–512. IEEE, 2019.

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

[148] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3):21:1–21:42, 2019.

[149] Tracy Jia Shen, Robert Cowell, Aditi Gupta, Thai Le, Amulya Yadav, and Dongwon Lee. How gullible are you?: Predicting susceptibility to fake news. In *WebSci*, pages 287–288. ACM, 2019.

[150] Shirish K. Shevade, S. Sathiya Keerthi, Chiranjib Bhattacharyya, and K. R. K. Murthy. Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Networks Learn. Syst.*, 11(5):1188–1193, 2000.

[151] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.

[152] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36, 2017.

[153] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *WSDM*, pages 312–320. ACM, 2019.

[154] Kamaldeep Singh, Sharath Chandra Guntuku, Abhishek Thakur, and Chittaranjan Hota. Big data analytics framework for peer-to-peer botnet detection using random forests. *Inf. Sci.*, 278:488–497, 2014.

[155] Neharika Singh and Madhumita Chatterjee. Botdefender: A framework to detect bots in online social media. *JNCET*, 7(9), 2017.

[156] Marshall Sponder and Gohar F. Khan. *Digital analytics for marketing.* Routledge, 2017.

[157] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. Social media analytics - challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.*, 39:156–168, 2018.

[158] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[159] Pablo Suárez-Serrato, Margaret E. Roberts, Clayton A. Davis, and Filippo Menczer. On the influence of social bots in online protests - preliminary findings of a mexican case study. In *SocInfo (2)*, volume 10047 of *Lecture Notes in Computer Science*, pages 269–278, 2016.

[160] V. S. Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The DARPA twitter bot challenge. *IEEE Computer*, 49(6):38–46, 2016.

[161] Kalpathy Ramaiyer Subramanian. Influence of social media in interpersonal communication. *IJSPR*, 38(2):70–75, 2017.

[162] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction.* Adaptive computation and machine learning. MIT Press, 2018.

[163] Bela Szunyogh. *Psychological warfare: An introduction to ideological propaganda and the techniques of psychological warfare*, volume 50. William Frederick Press, 1956.

[164] Fadi A. Thabtah, Peter I. Cowling, and Yonghong Peng. MMAC: A new multi-class, multi-label associative classification approach. In *ICDM*, pages 217–224. IEEE Computer Society, 2004.

[165] Ivan Tomek. Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.*, SMC-6(11):769–772, 1976.

[166] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM*, pages 280–289. AAAI Press, 2017.

[167] Monika Verma and Sanjeev Sofat. Techniques to detect spammers in twitter-a survey. *IJCA*, 85(10), 2014.

[168] Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information - a survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 7(5), 2017.

[169] Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In *#MSM*, volume 838 of *CEUR Workshop Proceedings*, pages 41–48. CEUR-WS.org, 2012.

[170] Randall Wald, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. Predicting susceptibility to social bots on twitter. In *IRI*, pages 6–13. IEEE Computer Society, 2013.

[171] Tyler Wall. Us psychological warfare and civilian targeting. *Peace Rev.*, 22(3):288–294, 2010.

[172] Binghui Wang, Neil Zhenqiang Gong, and Hao Fu. GANG: detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In *ICDM*, pages 465–474. IEEE Computer Society, 2017.

[173] Claire Wardle. Fake news. it's complicated. *First Draft*, 16:1–11, 2017.

[174] Claire Wardle and Hossein Derakhshan. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Journalism, 'fake news' & disinformation*, pages 43–54, 2018.

[175] Geoffrey I. Webb. Decision tree grafting from the all tests but one partition. In *IJCAI*, pages 702–707. Morgan Kaufmann, 1999.

[176] David Westerman, Patric R. Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *J. Computer-Mediated Communication*, 19(2):171–183, 2014.

[177] Philipp Wicke and Marianna M. Bolognesi. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *arXiv preprint arXiv:2004.06986*, 2020.

[178] Brenda K. Wiederhold. Social Media Use during Social Distancing, 2020.

[179] Christine B. Williams. Introduction: Social media, political marketing and the 2016 u.s. election. *J. Political Mark.*, 16(3-4):207–211, 2017.

[180] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Clim. Res.*, 30(1):79–82, 2005.

[181] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[182] Liang Wu, Fred Morstatter, Xia Hu, and Huan Liu. Mining misinformation in social media. In *Big Data in Complex and Social Networks*, pages 135–162. Chapman and Hall/CRC, 2016.

[183] Kai-Cheng Yang, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[184] Shuhan Yuan, Xintao Wu, Jun Li, and Aidong Lu. Spectrum-based deep neural networks for fraud detection. In *CIKM*, pages 2419–2422. ACM, 2017.

[185] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[186] Daniel Zeng, Hsinchun Chen, Robert F. Lusch, and Shu-Hsing Li. Social media analytics and intelligence. *IEEE Intell. Syst.*, 25(6):13–16, 2010.

[187] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)*, 53(5):1–40, 2020.

[188] Mingzhu Zhu, Chao Xu, and Yi-fang Brook Wu. IFME: information filtering by multiple examples with under-sampling in a digital library environment. In *JCDL*, pages 107–110. ACM, 2013.

[189] Arkaitz Zubiaga, Bahareh R. Heravi, Jisun An, and Haewoon Kwak. Social media mining for journalism. *Online Inf. Rev.*, 43(1):2–6, 2019.