U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

PhD
3.º CICLO
FCUP
2022

U.PORTO

Towards realistic scenarios concerning the identification of unreliable information in social networks

Nuno Ricardo Pinheiro da Silva Guimarães

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Towards realistic scenarios concerning the identification of unreliable information in social networks

Nuno Ricardo Pinheiro da Silva Guimarães

Tese de Doutoramento apresentada à
Faculdade de Ciências da Universidade do Porto
Ciência de Computadores

2022

PhD

Towards realistic scenarios concerning the identification of unreliable information in social networks

Nuno Ricardo Pinheiro da Silva Guimarães
Doutoramento em Ciência de Computadores
Departamento de Ciência de Computadores
2022

Orientador
Álvaro Pedro de Barros Borges Reis Figueira, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientador
Luís Fernando Rainho Alves Torgo, Professor Associado, Faculty of Computer Science Dalhousie University

Nuno Ricardo Pinheiro da Silva Guimarães

Towards realistic scenarios concerning the identification of unreliable information in social networks

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
2022

Nuno Ricardo Pinheiro da Silva Guimarães

Towards realistic scenarios concerning the identification of unreliable information in social networks

Thesis submitted to
Faculdade de Ciências, Universidade do Porto
to obtain the degree of Doctor in Computer Science

Supervisor: Prof. Álvaro Figueira
Co-supervisor: Prof. Luís Torgo

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
2022

To my parents

Acknowledgments

First and foremost I would like to highlight that during these unprecedented times we live in and with a global pandemic approaching its third year, this section carries a greater meaning and importance than it did at the beginning of my Ph.D. Hence, I would like to apologize for being this long but I do believe that the current situation justifies it.

Therefore, I would like to start by thanking my supervisors for their guidance through this journey. I am grateful for the freedom they allow me to have during this work and even though there were some "bumps" in the way, they were always supportive of the path that I took. I would like to express my gratitude to Prof. Álvaro Figueira. We have been working together since 2016 and his guidance and encouragement have helped me to grow as a researcher and a person. He encouraged me to submit a Ph.D. proposal for funding and without his initiative, support, and interest, this work would likely not exist. I am deeply grateful for his mentoring, trust, and words of support even when things didn't turn out so good. I also like to thank Prof. Luís Torgo for embarking on this adventure. Prof. Luís Torgo was the supervisor of my MSc thesis and I was thrilled when he accepted to join as a co-supervisor in my Ph.D. Even though we were in different countries, I never felt that his guidance was compromised and I knew that if I needed some help he would provide it. I would also like to thank the opportunity to work in a different country and the support he provided during my stay. It was an experience I will never forget.

I would like to take the opportunity to thank my colleagues/lunch buddies at Dalhousie University Ines, Lucas, Fateha, Varshid, Marzie, Luiz, and Alessandra for welcoming me into their group and taking me along in their adventures. In particular, I would like to thank Paula Branco for the opportunity to enroll in a posterior visiting research student program at the University of Ottawa and for the collaboration that resulted from it. In that regard, a special thanks also to Xuanyu Su for collaborating with me in the last steps of this work.

A word of gratitude goes for my colleagues Daniel Loureiro, Vitor Cerqueira, Shamsuddeen Muhammad, David Aparício, Vanessa Silva, Muhammad Abubakar Sadiq, and Mariana Oliveira whose conversations and feedback helped shape some parts of this work. A special

thanks to Nuno Moniz for the conversations throughout the years and the input and feedback given regarding my Ph.D. proposal. His feedback helped to secure funding for this work and thus I am very grateful for that. I would also like to thank Fundação para a Ciência e Tecnologia (FCT), Portugal for the Ph.D. Grant (SFRH/BD/129708/2017) and for extending the financial support during the pandemic that allowed the conclusion of this work. A special thank you goes also to Alexandra Ferreira, Isabel Gonçalves, and Isabel Paulo. Since I started as an undergraduate student in 2009 they were always willing to help no matter how small the problem was. To my friends Marcelo Santos, Joana Santos, Filipe Santos, and João Queirós, thank you for all the coffee breaks, lunches, and dinners we had. They were an essential distraction from work and helped me immensely. I also want to thank Cláudia Costa for her support and help since the beginning of my Ph.D..

A very special thanks to Joana Dumas. Her support and company in the last years of this work were essential and I will deeply miss our coffee/lunch breaks. Her friendship was one of the best things that happened during my Ph.D. I also want to thank Joana Gonçalves, Rui Fonseca, and Patrícia Santos for their friendship and for always being present all these years. Even with all the uncertainty and isolation during Covid lockdowns, a video call with them would always cheer me up. I would also like to express my gratitude to my "solar" friends António Pinto and Diogo Teixeira. Together, we achieved some things I would never dream of and their resilience and perseverance always inspired me in the course of this work. Similarly, a word of gratitude goes to Rui Gomes and Rafael Carolo. Thank you for all the laughs but also for always having some words of comfort or advice in less cheerful times. I've learned valuable lessons from both of you.

A special thanks to Tânia Rodrigues. She has been an incredible friend that helped me to keep my sanity intact during the pandemic. Her support and advice were essential for the completion of this thesis and her words always helped me to push through, one pomodoro at the time. I am very grateful to have her as a friend. Finally, to Jorge Silva, a very special thank you. From throwing pencil cases through the window 15 years ago to finishing a Ph.D. What a ride it was! All these years he was always an older brother to me: someone you look up to and is always there for you, but also someone who helped me professionally by being continuously available to discuss multiple ideas (always with a bit of "constructive criticism") during this work. He helped me to be a better researcher and a better person by leading by example and I am deeply grateful for that.

Finally, I would like to thank my family. They were always supportive of the path I took. In particular, I would like to thank:

My grandparents Guimas and Té whose kindness and long lunches and dinners I deeply miss.

My uncle José for his support and providing me a refuge where I could rest on multiple occasions during the course of this work.

My grandparents Albino and Luísa that continuously supported my education and made sure I would get this far.

My father Jorge. I am sure he would be immensely proud of what I achieved.

My mother Paula. I would like to take this opportunity to take back what I said 18 years ago when she was doing her Ph.D. It is definitely not an easy task! But it was also thanks to her that I decided to pursue it. She has been a constant inspiration in my life and I am sincerely grateful for all the things she did for me. She is the reason why I am here today. Thank you. This work is for you.

Abstract

In the last decade, social networks have evolved from platforms focused on connecting friends to include a fast-moving news medium where information reaches a global audience instantly. However, the anonymity provided to the users and the lack of gate-keeping concerning the content shared, have allowed malicious agents to spread unreliable information within the network, influencing users' perceptions and opinions on important issues. Due to the enormous amount of content produced on social media, there is a need for more automated methods to identify unreliable information and the accounts that spread it. Research on these two problems has been thoroughly conducted in recent years, but there is still a gap between experimental settings and real-world applications. In other words, it is still noticeable that the majority of studies lack a more pragmatic approach limiting the applicability of solutions in a realistic scenario. The work presented in this thesis attempts to bridge the gap between more experimental and realistic approaches in the detection of unreliable content and the accounts that distribute it. In the identification of accounts that distribute unreliable content, the bot detection task has already been studied in depth. However, not all bot accounts operate as distributors of unreliable content (e.g., news aggregation bots), and not all human accounts are necessarily reliable, with growing evidence suggesting that they play an important role in the propagation of unreliable content. Thus, we focus on distinguishing between unreliable and reliable accounts, regardless of how they are operated. We propose and assess the usefulness of knowledge-based metrics to evaluate accounts based on their impact and behavior. In addition, we also work towards a prediction-based detection system capable of dealing with real-world situations by introducing constraints on the content available on each account (based on volume- and time-based batches). Experiments conducted on a validation set with a different number of tweets per account provide evidence that solutions that adapt to the number of publications of each account lead to a performance improvement of up to 20% compared to traditional (individual) models and to cross-batch models (which perform better on different batches of tweets).

Regarding the automatic identification of unreliable/reliable content, current research on this topic has focused on specific contexts or events (such as elections). Thus, it is not clear whether the provided features and models can be effectively used in a real-world

application where the topics of reliable and unreliable content may change over time. Hence our contribution to this task is to longitudinally evaluate the current proposals in a long-term scenario using social network publications spanning an 18-months period. We experimented with 3 different scenarios where feature/model combinations are trained with 15-, 30-, 60-day data and evaluated over the remaining time period. Results show that detection models trained with word-embedding features (especially derived from BERT-based language models) perform better and are less likely to be affected by topic changes (e.g., the rise of Covid-19 conspiracy theories).

The results presented in this work provide the basis for more pragmatic approaches to the problems of detecting unreliable accounts and content in social networks, and bridge the gap between more experimental studies and real-world applications in this domain.

Resumo

Na última década, as redes sociais evoluíram de plataformas orientadas ao estabelecimento de ligações entre amigos até um meio de propagação de informação a alta velocidade onde a informação atinge uma audiência global de forma instantânea. Porém, a anonimidade fornecida aos utilizadores, assim como a falta de *gate-keeping* sobre o conteúdo publicado e partilhado têm permitido a difusão de conteúdo não-confiável pelas redes sociais com impacto na perceção e na opinião dos utilizadores em temas relevantes. Devido à elevada quantidade de conteúdo produzido, existe uma necessidade de métodos automáticos para identificação de informação não confiável assim como das contas que a propagam. Nos últimos anos a investigação científica nestes dois problemas tem sido realizada de forma intensa, permanecendo, no entanto, uma elevada discrepância entre as experiências conduzidas em contexto de investigação e sua aplicabilidade no "mundo real". Por outras palavras, é ainda notório que a maioria dos estudos carece de uma abordagem mais pragmática, sendo a aplicabilidade destas soluções, na prática, ainda bastante limitada. O trabalho apresentado nesta tese foca-se na aproximação de abordagens mais experimentais ao contexto do mundo real na deteção de conteúdo não confiável e das contas que o propagam. Na identificação das contas que distribuem conteúdo não confiável, a deteção de bots tem sido amplamente estudada na literatura. No entanto, nem todos os bots distribuem conteúdo não confiável (como por exemplo, bots que funcionam como agregadores de notícias), assim como nem todas as contas operadas por humanos são confiáveis visto que uma crescente evidência científica sugere um papel significativo destas na propagação de conteúdo não confiável. Desta forma, focamo-nos na distinção entre contas confiáveis e não confiáveis, independentemente do seu grau de automatização. Propomos e analisamos a utilidade de métricas baseadas em conhecimento prévio para a avaliação de contas com base no seu comportamento e impacto. Adicionalmente, propomos um sistema de base preditiva para deteção de contas não confiáveis, capaz de lidar com cenários do mundo real. Para esse efeito aplicamos restrições no conteúdo disponível em cada conta, impondo limitações por volume de posts ou posts limitados a um intervalo de tempo específico. A avaliação das soluções propostas num dataset de validação, em que os tweets por cada conta vão variando, mostra uma melhoria dos resultados até 20% comparativamente com soluções mais tradicionais, e contra soluções individuais que funcionam melhor, em média, nas diferentes restrições

impostas.

Relativamente à identificação automática de conteúdo, a literatura atual está direcionada para eventos ou contextos específicos onde informação não confiável é propagada (tal como eleições). Desta forma, não é claro que as *features* e os modelos propostos nestes contextos possam ser aplicados eficazmente num cenário próximo ao de um "mundo real", onde os tópicos discutidos no conteúdo confiável e não confiável podem mudar ao longo do tempo. Portanto, a nossa contribuição nesta tarefa consiste na avaliação longitudinal das diferentes propostas na literatura usando dados publicados nas redes sociais referente a um período de 18 meses. Foram considerados dados relativos a três intervalos de tempo (15, 30 e 60 dias) para o treino das diferentes combinações de *features* e modelos, sendo que os restantes foram usados para avaliação dos mesmos. Os resultados demonstram que modelos de deteção de conteúdo não confiável, treinados com *features* derivadas de *word-embeddings* (particularmente derivados dos modelos BERT) apresentam uma melhor performance e são menos afetados pela mudança de tópicos (por exemplo, com o surgimento de teorias de conspiração relacionadas com o Covid-19).

Os resultados apresentados neste trabalho providenciam os alicerces necessários para metodologias mais pragmáticas na deteção automática de contas e de conteúdo não confiável em redes sociais, aproximando os estudos mais experimentais a soluções capazes de serem implementadas no mundo real.

Contents

List of Tables **20**

List of Figures **25**

1 Introduction **27**

1.1 Historical Perspective 28

1.2 Contributions 30

1.3 Thesis Outline 31

1.4 Bibliographic Note 32

2 Literature Review **35**

2.1 Characterization of Unreliable Content 35

2.2 Malicious Actors in Social Networks 38

2.2.1 Non-Human Accounts 38

2.2.2 Human Accounts 39

2.3 Data Sources, Data Annotation and Datasets 40

2.3.1 Data Sources 40

2.3.2 Fact-Checking Sources 42

2.3.3 Data Annotation Methods 43

2.3.3.1 Human Annotation 43

2.3.3.2 Distant annotation 44

2.3.4 Datasets . . . 44
2.3.4.1 Dataset Comparison . . . 48
2.4 Research Paths . . . 49
2.4.1 Exploratory Analysis of Unreliable Information . . . 50
2.4.1.1 Case Studies . . . 51
2.4.2 Detection of Unreliable Content and Accounts . . . 53
2.4.2.1 Input Features . . . 54
2.4.2.2 Model Types . . . 57
2.4.2.3 Evaluation Metrics . . . 59
2.4.3 Network Based Solutions . . . 62
2.5 Discussion . . . 64

3 Data Extraction and Preliminary Exploratory Analysis 67

3.1 Data Extraction . . . 67
3.1.1 Account Annotation . . . 72
3.1.2 Datasets . . . 73
3.2 Exploratory Analysis . . . 73
3.2.1 Longitudinal Analysis . . . 74
3.2.2 Account Analysis . . . 80
3.2.3 Comparative Analysis . . . 83
3.3 Conclusions . . . 89

4 Towards a Pragmatic Detection of Unreliable Accounts 91

4.1 Knowledge-Based Approach . . . 92
4.1.1 Case Studies . . . 96
4.1.1.1 Top Unreliable and Reliable Accounts . . . 97
4.1.1.2 Botometer vs Reputation Metrics . . . 100

4.1.1.3 Twitter Unreliable Accounts Evaluation 102
4.1.2 Conclusions 106
4.2 Classification-Based Approach 108
4.2.1 Experimental Setup 109
4.2.1.1 Data Extraction 109
4.2.1.2 Topic and Keyword Analysis 110
4.2.2 Tweets Batch Size and Time Intervals 112
4.2.3 Feature Extraction 112
4.2.3.1 Account-based features 113
4.2.3.2 Content-based features 113
4.2.4 Feature Selection 113
4.2.5 Model Training and Evaluation 114
4.2.6 Validation 114
4.2.7 Results 116
4.2.7.1 Feature Analysis 116
4.2.7.2 Model Evaluation 117
4.2.7.3 Validation in Random Batches 124
4.2.8 Conclusions 125
4.3 Chapter Summary 126

5 Performance of Unreliable Detection Models in Twitter Posts Over Time 129

5.1 Problem Statement and Proposed Solution 129
5.2 Experiment Workflow 131
5.2.1 Data Extraction 131
5.2.2 Feature Extraction 132
5.2.3 Feature Analysis 134
5.2.4 Models and Evaluation 134

5.3 Results 134
5.3.1 Feature Importance 134
5.3.2 Models' Evaluation 138
5.4 Performance Boosting with BERT and RoBERTa Derived Features 142
5.5 Conclusions 146

6 Conclusions **149**
6.1 Strengths and Limitations 150
6.2 Future Work 153
6.3 Final Remarks 155

A Feature Importance **157**

B Models Performance **161**
B.1 Models Performance Using BERT and RoBERTa Features 161

References **166**

List of Tables

2.1 Concepts of unreliable information 37

2.2 List of relevant works grouped by data source. 42

2.3 State of the art datasets 49

2.4 Case studies in unreliable information 53

2.5 Example of a confusion matrix for binary classification. 60

3.1 Distribution of websites per unreliable class in OpenSources. 69

3.2 Distribution of websites per reliable class in MBFC. 70

3.3 Entities frequency in reliable and unreliable tweets 88

4.1 Characterization of the accounts with the highest Botometer score 101

4.2 Characterization of the accounts with the highest IMP score 101

4.3 $PCOUNT$ scores in the datasets extracted from the Twitter API 104

4.4 BEH and BEH_SF scores in the datasets extracted from the Twitter API . 105

4.5 IMP and IMP_SF scores in the datasets extracted from the Twitter API . 105

4.6 Topics discussed in reliable and unreliable accounts 111

4.7 Feature importance scores for each batch 118

4.8 Cross-batch results for volume-based batches 122

a 5 features 122

b 10 features 122

c 15 features 122

d 20 features 122
e 25 features 122
f 30 features 122
g 35 features 122
4.9 Cross-batch results for time-based batches 123
a 5 features 123
b 10 features 123
c 15 features 123
d 20 features 123
e 25 features 123
f 30 features 123
g 35 features 123
4.10 Results of the different models in the validation dataset 125

5.1 Feature importance in the 15-, 30-, and 60-day windows 138
5.2 Main differences between the 4 BERT-derived models 143

List of Figures

1.1 Number of hits per year in Google Scholar for the term "fake news" 29

2.1 Diagram of the main concepts of unreliable information 38

2.2 Diagram of the main data sources of unreliable information 42

2.3 Diagram of the main data annotation methods presented in the current literature regarding unreliable information. 44

2.4 Social media components 55

a Account profile 55

b Post information 55

c Network Information 55

3.1 Data extraction workflow 71

3.2 Examples of tweets extracted 72

3.3 Volume of tweets, unique sources and unique classifications aggregated by month. 74

3.4 Negative and positive sentiment scores of unreliable tweet averaged by month. 76

3.5 Emotions scores of unreliable tweet averaged by month 77

3.6 Wordclouds for the top 20 most frequently used hashtags in unreliable tweets captured by month. 78

3.7 Wordclouds for the top 20 most frequently used entities in unreliable tweets captured by month 79

3.8 The number of posts discriminated by type of unreliable content for the top 30 accounts. 81

3.9 Number of accounts that diffuse content in each pair of classes. 82

3.10 Unreliable accounts' status at the time of verification (November 2021). . . . 83

3.11 Sentiment score for reliable and unreliable tweets aggregated by day 85

3.12 Emotions scores for reliable and unreliable tweets aggregated by day 86

3.13 Wordclouds of entities and hashtags regarding reliable and unreliable tweets. 87

4.1 Experimental system for account reliability classification 91

4.2 Characteristics of the accounts with the highest IMP_{SF} 98

4.3 Comparison between IMP and Botometer score for the accounts retrieved . . 100

4.4 Diagram of the experimental setup 109

4.5 Wordclouds for the most relevant keywords in each class 112

a Reliable wordcloud 112

b Unreliable wordcloud 112

4.6 Performance variation for the different models and number of features using account-only data (tweets=0). 119

4.7 Performance of the different models with the variation of the volume batches in the selected features set 120

4.8 Performance of the different models with the variation of the time batches in the selected features set 121

5.1 The three different scenarios considered concerning the size of the training set 132

5.2 Importance of 30 most relevant features from different feature sets using a 15-day interval 136

a Bag of Words 136

b Google Word2Vec 136

c Fake Word2Vec 136

d Context-free features 136

e Lexical Categories (Empath) 136

5.3 Performance evaluation of different models for each feature set 140

a Bag of Words . . . 140
b Google Word2Vec . . . 140
c Fake Word2Vec . . . 140
d Lexical Categories (Empath) . . . 140
e Context-free features . . . 140
f Best 15 features . . . 140
5.4 Comparison of the best models for each scenario using Google pre-trained Word2Vec . . . 141
5.5 Comparison of best models for each scenario using fake Word2Vec . . . 141
5.6 Distribution of tokens in tweets . . . 143
5.7 Performance evaluation (using weighted F1-score) of the different models using BERT-derived features . . . 144
a Google Word2Vec . . . 144
b Fake Word2Vec . . . 144
c BERT-base . . . 144
d BERT-large . . . 144
e RoBERTa-base . . . 144
f RoBERTa-large . . . 144
5.8 Comparison of the best models for each scenario using RoBERTa-base features 145
5.9 Comparison of the best models for each scenario using RoBERTa-large features 146
A.1 Importance of the 30 most relevant features in 30-day interval . . . 158
a Bag of Words . . . 158
b Google Word2Vec . . . 158
c Fake Word2Vec . . . 158
d Context-free features . . . 158
e Lexical Categories (Empath) . . . 158
A.2 Importance of the 30 most relevant features in 60-day interval . . . 159

a Bag of Words . . . 159
b Google Word2Vec . . . 159
c Fake Word2Vec . . . 159
d Context-free features . . . 159
e Lexical Categories (Empath) . . . 159

B.1 Performance evaluation of the different models and features in 30-day training data . . . 162
a Bag of Words . . . 162
b Google Word2Vec . . . 162
c Fake Word2Vec . . . 162
d Lexical Categories (Empath) . . . 162
e Context-free features . . . 162
f Best 15 features . . . 162

B.2 Performance evaluation of the different models and features in 60-day training data . . . 163
a Bag of Words . . . 163
b Google Word2Vec . . . 163
c Fake Word2Vec . . . 163
d Lexical Categories (Empath) . . . 163
e Context-free features . . . 163
f Best 15 features . . . 163

B.3 Performance evaluation of the different models and BERT-derived features in 30-day training data . . . 164
a BERT-base . . . 164
b BERT-large . . . 164
c RoBERTa-base . . . 164
d RoBERTa-large . . . 164

B.4 Performance evaluation of the different models and BERT-derived features in 60-day training data 165

a BERT-base 165

b BERT-large 165

c RoBERTa-base 165

d RoBERTa-large 165

Chapter 1

Introduction

The rise of news media such as newspapers and later radio and television contributed towards a more informed and educated population. Indeed, news have supported a better understanding of the world and its events. Due to a large set of standards and rules of conduct that govern journalism (code of ethics), news are subject to scrutiny by different entities before being published or broadcast on radio and television. Although this code of ethics varies from country to country, there are common principles such as truthfulness, accuracy, objectivity and impartiality that must be observed in the majority of the countries with freedom of the press.

With the advent of the digital age, news media had to adapt to the growing popularity of the Internet. Social media in particular have revolutionized the way information is consumed and how socialization and communication are perceived. It is estimated that there are currently 3.8 billion social media users, using 14 different platforms with more than 300 million monthly active users [56]. These factors have led to an overwhelming amount of information available on a daily (or even hourly) basis. In addition, easy access to these platforms, whether through a computer, tablet, or mobile/cell phone, makes the consumption and dissemination of information almost instantaneous.

With the shift towards a global digital medium, social networks in particular have begun to function as platforms for information aggregation. Enterprises and important figures (such as politicians, celebrities, etc.) have begun to use social networks to reach their audiences. With the decline of traditional news media such as newspapers [189], news media entities have also engaged on social networks to reach their audiences. This changed the landscape of social networks from platforms where the information shared was personal and focused on relationships with friends to a news medium where breaking news and events are reported in a matter of seconds. In fact, a 2018 study concluded that 68% of American adults use social media (at least occasionally) for their daily news consumption [204]. Another study

also concluded that in 2019 40% of internet users adopt social media to stay up-to-date as news is concerned. In 2020 and 2021 more than half of Twitter users regularly used the social network as a news medium [248].

However, some characteristics of social networks such as the ability to create accounts with ease and a high degree of anonymity enable the intrusion of bad actors whose main goal is to destabilize the social network ecosystem. This destabilization can be achieved, for example, through the distribution of spam content (similar to what happens in e-mails), malware in the form of links, and unreliable information, the latter being the main topic discussed in this work.

1.1 Historical Perspective

The rise of news consumption in social media, the ease of creating and disseminating content, and the lack of control mechanisms on these platforms have led to an increase in unreliable content. However, the dissemination of unreliable content is not a recent problem and has been present in society in various forms. Although the term gained popularity during the 2016 presidential election in the United States, there have been several examples of the spread of unreliable information over the years that have had serious consequences in the real world. For example, in 1924 a forged document known as "The Zinoviev Letter" was published in a well-known British newspaper four days before the general election. Its aim was to destabilize the elections in favor of the Conservative Party, with a directive from Moscow to the British Communists referring to an Anglo-Soviet treaty and encouraging "agitational propaganda" in the armed forces [162]. Another example occurred after the "Hillsborough accident", in which 96 people were crushed to death because of overcrowding and inadequate security. Reports in an illustrious newspaper alleged that, as people were dying, some fellow drunken supporters stole from people and beat police officers who tried to help. However, these allegations later proved to be false [50].

Before online social networks existed, one of the pieces of false information that probably had a major impact on modern history was the claim that HIV was fabricated in a facility in the United States [28]. The spread of this rumor was an attempt by the KGB during the Cold War to sabotage the United States credibility. The false information began circulating in 1983 and was later picked up by an American television station [65]. Although the rumor was later debunked, the consequences are still present today with some studies suggesting the existence of a high percentage of people that believe in HIV-related hoaxes [27, 129].

A similar approach was taken in 2016. An investigation conducted by the Special Counsel's office (commonly known as "The Muller Report") provided evidence of Russian interference in the United States presidential election [157]. According to the report, the Russian opera-

tion began in early 2014 when social media pages and accounts were created to attract American audiences by members of the Russian organization IRA (Internet Research Agency). Two years later, these pages disseminated content supporting Republican candidate Donald Trump and defamatory content towards Democratic candidate Hillary Clinton. Related to the same event, an investigation led by Buzzfeed also concluded that more than 100 websites containing false information about both candidates were operated by a small group of teenagers in Macedonia. [215].

Recently, unreliable information about the global pandemic caused by the coronavirus (Covid-19) has been disseminated. This has shaped people's behavior toward the pandemic, neglecting basic recommendations promoted by trusted health organizations. In addition, unreliable information about vaccines also had an impact on the intent of vaccination in the United States and the United Kingdom [142].

The impact that unreliable information had on the 2016 presidential campaign in the United States led to the coining of the term "fake news" and highlighted the importance and consequences of this problem when information is disseminated in a fast-paced medium such as social networks. As a result, technology companies such as Google, Facebook and Twitter are working on various solutions to mitigate this problem [105, 107].

The scientific community has also been addressing this issue. Figure 1.1 shows the number of hits per year in Google Scholar regarding the term "fake news", where we can observe a constant growth in the number of publications on this topic. In particular, in 2017 there was an exponential growth in the number of publications (approximately 9300) compared to the previous year.

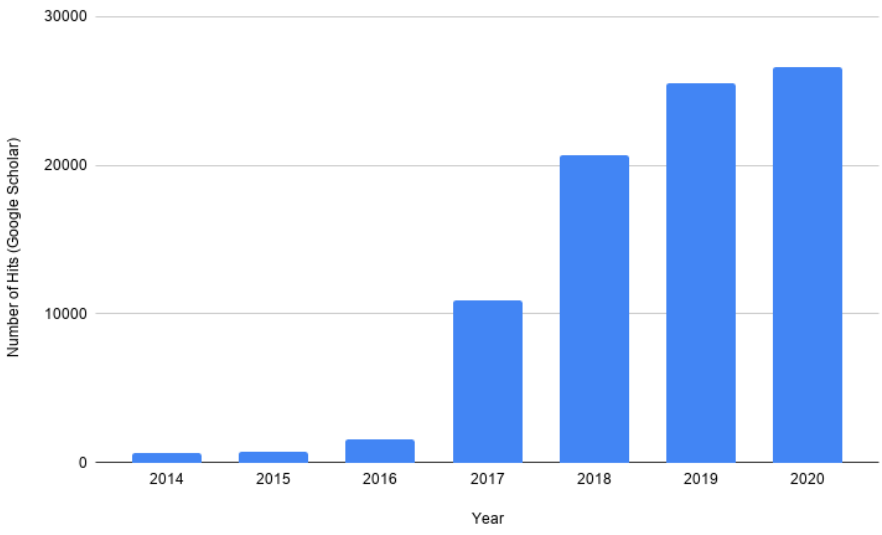


Figure 1.1: Number of hits per year in Google Scholar for the term "fake news"

1.2 Contributions

In social networks like Twitter, the problem of debunking and limiting the spread of misinformation is complex. First, with over 330 million monthly active users [219], the amount of content generated far exceeds the ability to analyze it manually. Furthermore, it is not trivial to distinguish accounts that publish/propagate unreliable content from accounts that do not.

These two problems led to the use of data mining and machine learning techniques to address the issue. First, data mining provides tools to analyze large amounts of data and extract meaningful knowledge that would otherwise be ignored. Second, this knowledge can be used to develop indicators and create various machine learning models/systems that can help detect unreliable content and malicious accounts in social media.

This thesis primarily focuses on using these data mining and machine learning techniques to address two different problems concerning unreliable information in social media. The first is the detection of unreliable accounts on social media (i.e. accounts that post and share unreliable content) and the second is the detection of unreliable content/posts.

While there has been extensive research on these two problems, the limitations imposed by real-world scenarios create a gap between experimental approaches proposed and more fully developed solutions. In the detection of unreliable accounts, current research has focused on the development of bot detection systems. Although these systems limit the dissemination of unreliable information, they do not capture the entirety of accounts that disseminate this type of content, as human accounts are also largely responsible for disseminating unreliable content [247]. Moreover, most approaches disregard the number of posts available from each account at the time of detection and do not consider this factor in the experimental design. This may limit the effectiveness of such approaches in more realistic scenarios, as the number of posts on social media accounts may vary significantly.

Regarding the detection of unreliable posts on social media, current approaches target specific events or small time intervals and often ignore the temporal dependency that might be associated with the task. In a more pragmatic scenario, varying the style or topics of unreliable content can affect the performance of detection models.

In this thesis, we address the problems mentioned earlier. First, we investigate if the bot detection problem needs to be addressed comprehensively to account for the inclusion of unreliable human accounts. Second, we evaluate the behavior of current features and models used in unreliable information detection and similar tasks, and test their longevity over time.

The work presented in this thesis can be summarized in the following contributions:

1. a comprehensive review of the current literature is provided, highlighting some caveats and open questions regarding unreliable accounts and content on social networks;

2. a distance labeling extraction methodology is adapted from the state of the art to retrieve unreliable and reliable tweets and accounts; in addition, a 3-step preliminary exploratory analysis of the data is conducted regarding the characteristics of unreliable content over time, unreliable accounts, and similarities and differences between reliable and unreliable posts

3. knowledge-based metrics are proposed to leverage the information extracted by the aforementioned methodology, and an analysis of the extracted accounts is performed with 3 different case studies;

4. we address the task of detecting unreliable accounts, which, unlike the majority of previous work, is not limited to distinguish bot and human-operated accounts, but unreliable and reliable accounts as a whole. Moreover, we approach the problem in a more pragmatic scenario assuming that accounts may differ in terms of volume and frequency of tweets published;

5. we focus on evaluating how models and features proposed in the current state of the art perform over time;

1.3 Thesis Outline

This thesis is organized in 6 chapters. A summary of each one is described below

- **Introduction**: Chapter 1 presents a brief introduction to the problem of unreliable content. In addition, the motivation and contributions of this work are described.

- **Literature Review**: in Chapter 2 an extensive review of the state of the art is detailed, with a specific focus on the data, features, models, and tasks usually targeted in the context of unreliable content detection in social media.

- **Data Extraction and Preliminary Exploratory Analysis**: Chapter 3 describes the data extraction methodology used in this work. In addition, a preliminary exploratory data analysis grouped into three categories is conducted: a longitudinal analysis focusing on the characteristics of unreliable content over time, an account-based analysis targeting accounts that diffuse unreliable content, and a comparison-based analysis addressing the differences and similarities between reliable and unreliable posts.

- **Towards a Pragmatic Detection of Unreliable Accounts**: Chapter 4 presents our approach towards the detection of unreliable accounts. Two approaches are presented: a knowledge-based approach that uses metrics to classify the impact of accounts in social networks and a supervised machine learning approach that focuses on a more pragmatic scenario where the volume of posts associated with each account is dynamic.
- **Performance of Unreliable Detection Models in Twitter Posts Over Time**: Chapter 5 describes our experiments in the evaluation of unreliable detection systems over time and how state of the art features and models behave in a scenario where the topics discussed can change
- **Conclusions**: Finally, Chapter 6 of this thesis presents the conclusions, strengths, and limitations of the current work. In addition, it illustrates possible research paths to close the gap between more experimental approaches to real-world scenarios in the detection of unreliable content and the accounts that diffuse it.

1.4 Bibliographic Note

The following list provides some of the work already published in the scope of this thesis.

- A. Figueira, N. Guimaraes, L. Torgo: Current State of the Art to Detect Fake News in Social Media: Global Trendings and Next Challenges in Proceedings of 14th International Conference on Web Information Systems and Technologies, 2018 (Chapter 2)
- A. Figueira, N. Guimaraes, L. Torgo: A Brief Overview on the Strategies to Fight Back the Spread of False Information in Journal of Web Engineering, Vol 18, Pages 319-352,June 2019 (Chapter 2)
- N. Guimaraes, A. Figueira, L. Torgo: Contributions to the Detection of Unreliable Twitter Accounts through Analysis of Content and Behaviour in Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2018 (Chapter 3)
- N. Guimaraes, A. Figueira, L. Torgo: Analysis and Detection of Unreliable Users in Twitter: Two Case Studies in Knowledge Discovery, Knowledge Engineering and Knowledge Management (10th International Joint Conference, IC3K 2018, Seville, Spain, September 18-20, 2018, Revised Selected Papers), 2020 (Chapter 3)
- N. Guimaraes, A. Figueira, L. Torgo: Knowledge-Based Reliability Metrics for Social Media Accounts in Proceedings of the 16th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST, 339-350, 2020 (Chapter 4)

- N. Guimaraes, A. Figueira, L. Torgo: Towards a pragmatic detection of unreliable accounts on social networks in Online Social Networks and Media - Volume 24, July 2021 (Chapter 4)
- N. Guimaraes, A. Figueira, L. Torgo: Can Fake News Detection Models Maintain the Performance through Time? A Longitudinal Evaluation of Twitter Publications in Mathematics 2021, 9, 2988. (Chapter 5)

In addition, two news articles (in Portuguese) were published in collaboration with journalist Micael Pereira containing partial data and analysis from this thesis in the context of the Covid-19 pandemic [169, 170].

Chapter 2

Literature Review

Unreliable information has been thoroughly studied in the last few years due to the problems and impact caused by the diffusion of this type of content online. Research on this domain has branched over different paths, from the analysis and exploration of different types of unreliable content in social media to the implementation and detection of solutions to mitigate its propagation. In this chapter, we present an overall description of this research domain, with a particular focus on the two main topics discussed in this work: the analysis and detection of unreliable content published in social media and the accounts that publish or disseminate it.

2.1 Characterization of Unreliable Content

To better understand the subsequent work in this thesis, it is important to define what is unreliable content. Although fake news is currently the "buzzword" in the literature, there is no clear consensus amongst researchers on its definition. For example, the authors in [211] define fake news as "intentionally written to mislead consumers" while in Potthast et al. [176], fake news is defined as "the observation that, in social media, a certain kind of 'news' spread much more successfully than others, and that these 'news' are typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths". Another example can be seen in [77] where the authors adopt the definition that fake news websites are those which "intentionally publish hoaxes and disinformation for purposes other than news satire". Disinformation is also a commonly used concept in the area and it was originally defined as "information that is spread originally to deceive" [1]. "Disinformation" and "fake news" are concepts that are often used interchangeably, with the first being associated with a more historical perspective and the latter to more recent events within a social media

[1] https://www.lexico.com/definition/disinformation

context. However, other similar concepts are important to mention and consider when discussing unreliable information.

The combination of social media platforms with ad-based revenue in websites (i.e. website owners being paid for each time a certain page with an ad is visited) led to the definition of the term clickbait. This concept can be described as content published on the internet to draw users' attention and encourage them to click a link or go to a certain webpage [2]. From a social media perspective, this can be seen as a post with an engaging title leading to a full article that does not live up to the readers' expectations [41]. Similar to the previous concepts, clickbait can be associated with unreliable content since the headline is often not representative of the content of the article.

Two other concepts related to unreliable content are hate and extremely biased information. It is important to highlight that hate-related information is different from hate speech. In the scope of this work and in related literature, the former is referred to as information or content that promotes discrimination such as racism and homophobia. Hate speech, however, can be seen as a broader domain encompassing not only information-based sources but also incitement to violence or hatred from a group of people toward others based on, for example, race, color, and religion. On the other hand, extremely biased content can be seen as extremely one-sided information, often highlighting one side of the story at the expense of other perspectives (for example by using inflammatory and emotional tones or omitting relevant aspects of the story). This concept is frequently used in a political context and is often referred as hyper-partisan news [176]. However, it is important to emphasize that bias in news occurs at reasonable levels in mainstream news outlets [151], but extreme bias refers to content with decontextualized information and opinions disguised as facts.

Another concept that arises when discussing unreliable content is junk science. This refers to the promotion of pseudo-science and dubious scientific claims and rumors. Topics commonly discussed in this domain include miraculous cures through the use of natural/appeal to nature fallacies (i.e. the argument that if something is natural then it is also good), climate change denial, and anti-vaccination propaganda.

The last two concepts we describe are conspiracy theories and rumors. The first can be defined as stories that attempt to justify or explain a situation/event by relying on arguments without proof. For example, simplifying the intricacies of the real world by portraying social and political events as plots conceived by powerful entities [23]. On the other hand, rumors can be defined as stories whose truthfulness is ambiguous or never defined. However, like "fake news", the concept of "rumor" is not consistent throughout the literature. For example, a rumor can be defined as a "declaration that is generally plausible, associated with news, and is widespread without checking" [233] or "controversial and fact-checkable statement"

[2] https://www.oxfordlearnersdictionaries.com/definition/english/clickbait

[120]. Some literature [210, 208] also addresses a specific type of rumor - gossip - which can be defined as celebrity-driven rumors.

The concepts discussed previously are those that better fit the definition of "unreliable" used in this thesis, as we will discuss later in Chapter 3. It is also important to note that some of these concepts may overlap in their definition and therefore are not completely independent. For example, clickbait can be incorporated with junk science information (e.g. "This special fruit can cure Alzheimer!!!"). There are, however, other concepts discussed in the literature that we will briefly mention. Satire represents exaggeration and ironic information, usually related to current events, which can be seen as unreliable information without the proper identification. However, due to the specificity of this concept and the fact that satire sources and posts are more likely to be recognized as such, we do not consider this type of information as unreliable in the present work. Misinformation can be defined as false information that is published/disseminated regardless of the intention to deceive. Thus, the definition of misinformation relies solely on the intent of the actor passing on the information, and not on the characteristics of the information itself.

There are several more similar concepts in the literature. However, in their majority, they overlap with the ones previously presented. Common examples include credibility assessment [38, 98, 202] , deceptive news [193], hoaxes [223, 225], and propaganda [72, 92, 164].

In summary, unreliable information and similar concepts often overlap in the literature or take on different meanings depending on the study conducted. Some works attempt to standardize such concepts. For example, Zhou et al. [266] propose two different concepts for fake news: a broader definition and a narrower definition, and present a comparison of the different concepts regarding authenticity, intent, and news content. Another work [193], divides the concept into 3 different categories. Our attempt to aggregate and summarize the main concepts presented in the literature and discussed in the current section is presented in Figure 2.1. In addition, Table 2.1 presents some relevant papers summarized according to the concepts previously discussed.

Table 2.1: Some of the concepts of unreliable information associated with some relevant works in each topic.

Concept	Examples of Publications
Bias	[178, 44]
Clickbait	[5, 41, 42]
Conspiracy	[23, 240, 200]
Misinformation	[203, 138, 255]
Fake News	[171, 211, 30, 232, 213]
Rumour	[250, 269, 137, 63]

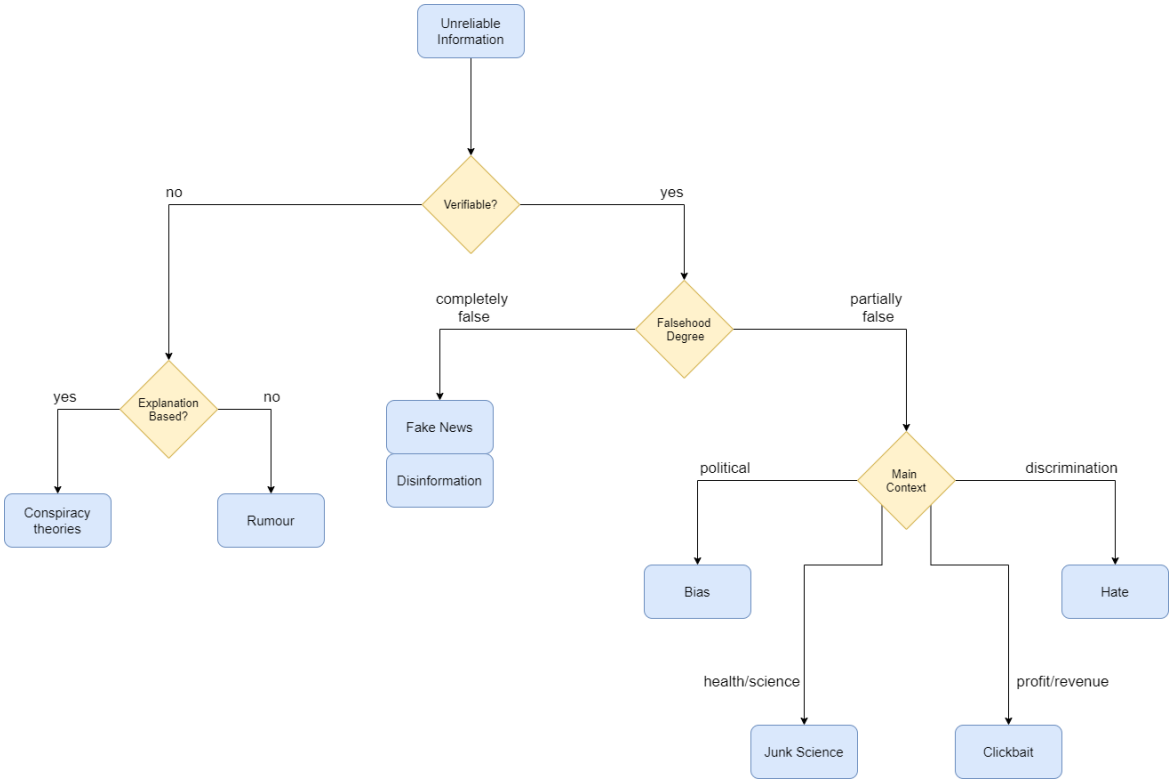


Figure 2.1: Diagram of the main concepts presented in the current literature regarding unreliable information.

2.2 Malicious Actors in Social Networks

In addition to the websites that publish unreliable content online, a key factor for the dissemination of this content can be attributed to malicious actors on social networks.

Malicious actors are responsible for publishing and spreading posts that pollute the social network ecosystem such as unsolicited content, phishing, and unreliable information.

These malicious actors can be divided in two different categories: non-human and human-operated accounts.

2.2.1 Non-Human Accounts

In the non-human category, the majority of the works in the literature has dealt with the analysis and detection of social bots.

Social bots can be defined as accounts that are operated automatically. More specifically, social bots produce content and interact with other accounts (bots or non-bots) without any type of human intervention. By definition, social bots are not malicious (for example,

some have the goal of aggregating news). However, malicious social bots have the goal of modifying or influencing behavior, causing a major impact in real-world scenarios whether by shifting public opinion in elections or the stock market [75]. Since bots are very active at an early stage when an unreliable article is published, a bot classification system capable of a timely detection can be an efficient strategy to prevent the spread of unreliable content in the network [202, 203]. Bot detection studies are discussed later in this chapter.

2.2.2 Human Accounts

It is important to highlight that although social bots amplified discussion on social networks, it is the human-operated accounts that are largely responsible for the proliferation of bot-generated content [75, 127].

Human accounts that disseminate unreliable content can be classified mainly by their intent (i.e., either they are aware that the disseminated content is unreliable or not).

"Trolls" are an example of human-operated accounts that intentionally disseminate unreliable information because their main purpose is to disrupt online conversations with inflammatory or off-topic content, consequently provoking emotional responses from other human-operated accounts. Evidence also suggests that, in some cases, troll accounts use bot accounts to amplify their messages or content [14]. Another example is compromised accounts, which are legitimate accounts accessed and operated by an illegitimate user. Under these circumstances, these accounts can be used to spread unreliable information such as fake news [179]

However, benign human-operated accounts, although not intentionally, are the main ones responsible for the propagation of unreliable information online [247]. The reasons why these accounts engage in this type of content are manifold but the vast majority of them, have their basis in social science theories. We address some of the most important concepts to better understand the main motivations behind the diffusion of unreliable content by these accounts.

Echo chambers and Filter Bubbles Echo chambers can be defined as a particular situation in which certain ideas and beliefs are constantly reinforced in a close-knit group because people are constantly exposed to the same viewpoints and lack of opposing voices. In a social network scenario, echo chambers arise not only because users' tend to follow and be followed by accounts with similar views, but also by a similar phenomenon created by the constraints imposed by recommendation algorithms in these platforms, that model the content displayed according to users' individual preferences. The term filter bubble was coined to highlight this phenomenon. Filter bubbles limit the diversification of content in users'

feeds because they are tailored to each individual user. Consequently, they narrow opinions and ideas, with the information shared in these clusters having very similar perspectives, which reinforces phenomena like confirmation bias and the "bandwagon effect".

Confirmation Bias and Bandwagon Effect Confirmation bias is the phenomenon that drives individuals to seek, interpret, and trust information that is consistent with their pre-existing beliefs [187]. Within social networks and social media, the confirmation bias is only amplified by filter bubbles and echo chambers. In addition, users also tend to be persuaded when they are exposed to the same content from several different sources. This phenomenon is called "bandwagon effect" and is based on the principle that the probability that an individual endorses a certain idea increases as more others endorse it too [198].

All of these factors contribute towards an involuntary spreading of unreliable content by naive human-operated accounts. This highlights the importance of distinguishing not only bot from human accounts but also unreliable from reliable accounts, regardless of the automation involved.

2.3 Data Sources, Data Annotation and Datasets

In this section, we discuss the different sources of unreliable information available online as well as how the data is verified and annotated. We finalize this section with some of the datasets already presented in the state of the art.

2.3.1 Data Sources

When it comes to the origin of unreliable content data there are two main groups: websites whose content/articles contain unreliable information and social media where unreliable posts are created (in some cases, linking to the unreliable articles on the websites), disseminated, and discussed.

There are clear differences between these two groups. Articles published on websites are longer and may include different types of multimedia content such as images and videos. On the other hand, unreliable information presented on social media is often limited to short texts and the media associated with a particular publication is limited. Websites also have custom layouts and designs, while publications created on social media are standardized according to the platform. This distinction is important because early work in this area concluded that technical qualities such as user interface design, usability/accessibility, and interactivity of websites are important credibility indicators for readers [253]. In social media, content is more uniform and thus the majority of these types of indicators are lost since posts

containing reliable and unreliable information present similar layout and design.

Research in unreliable information on social media has explored different platforms where this type of content proliferates. To better understand the nature of these studies, we provide a brief description of the characteristics of the main platforms used in the current literature:

- **Twitter**: Twitter is a micro-blogging platform that allows users to create posts up to a limit of 380 characters. It offers commonly social network characteristics such as following and be followed by other accounts, liking/"favoriting", replying to, or sharing posts. As of October 2021, Twitter had an average of 436 millions monthly active users and is one of the most used social media platforms in the world [220]. Combined with the accessibility and ease of data retrieval via the Twitter API, this social network is one of the most important sources of data for studies of unreliable information in social media.

- **Facebook**: Facebook is the most used social media platform in the world with 2740 million active users. Facebook offers similar features to Twitter such as the ability to like or react, comment on, and share posts from other accounts. However, Facebook presents a more friend-based platform than Twitter, with the main goal of the social network being connections between friends and family while Twitter philosophy is more related to the connections of ideas and topics. As far as data accessibility goes, since the Cambridge Analytica data scandal [35], Facebook has imposed several restrictions to its API, limiting the access for researchers and consequently affecting the studies on this social network [15].

- **Weibo**: Is a microblogging social network with strong similarities to Twitter. However, it presents a more regional popularity being one of the most used social networks in China. Feature-wise Weibo presents several similarities with other social networks with users being able to comment, like, and share posts from other accounts. Sina Weibo also offers an OpenAPI which allows for simplified and easy way to extract data.

- **Reddit**: Reddit departs from the traditional social networking platform by offering a forum-based layout in which users can post content in sub-forums (each on a specific topic). However, since Reddit has approximately 430 million monthly active users and has an easy to access API, some studies have been conducted on this platform regarding the publication of unreliable information.

- **WhatsApp**: WhatsApp is an instant messaging platform. However, the ability to create groups of more than a hundred users has had an impact on how the application is perceived. WhatsApp's popularity is particularly noticeable in countries such as India and Brazil with 390 and 108 million monthly active users, respectively [117] . Consequently, it has been a major platform for sharing unreliable information in

Table 2.2: List of relevant works grouped by data source.

Data Source	Relevant works
Twitter	[43][104][89] [96][67][61][31][71][38][246][247]
Facebook	[34] [267] [223] [181] [58] [118] [23][196][60][143]
Weibo	[261][257][259][254][121][52]
WhatsApp	[81][148][190][36]
Reddit	[54][53][199]
Website Articles	[112][242][176][180][132]

these countries [86, 148]. Moreover, WhatsApp is a private communication application, where data is encrypted end-to-end. Therefore, no API is available making the research on this platform very limited.

There are other social media platforms where unreliable information propagates. Common examples are YouTube and Tumblr. However, the number of studies on these platforms is still small. Some examples conducted on these platforms include the work of Hussain et al. [115] in which the authors use metadata from YouTube videos to assess the behavior of users in comments as well as analyzing bot-related patterns. Concerning Tumblr, the work in [160] provides a case study on the behavior of a disinformation dissemination account.

Figure 2.2 provides a diagram that aggregates and summarizes the different data sources. Table 2.2 provides some relevant literature that use the aforementioned sources.

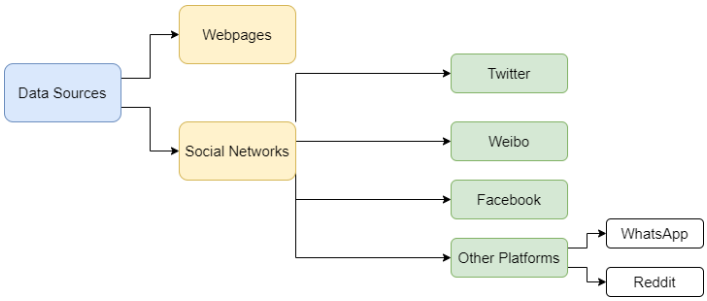


Figure 2.2: Diagram of the main data sources presented in the current literature regarding unreliable information.

2.3.2 Fact-Checking Sources

Several studies also rely on fact-checking websites to extract useful data to tackle the problem of unreliable information. These websites present claims that propagate in social media, annotated by experts in several degrees of truth. Several fact-checking websites exist, usually tackling claims on various topics. For example, Politifact covers claims more

related to American politics, while Snopes also includes other social issues such as health- and terrorism-related claims. In addition, several news media sources have developed their own fact-checking sections such as The Washington Post [3] and The New York Times [4]. Furthermore, country-specific allegations are being addressed by region-specific fact-checking initiatives (such as United Kingdom's Full Fact [5]), with Poynter Institute developing the International Fact Checking Network (IFCN) which reunites several fact-checking initiatives worldwide. A complete list of fact-checking websites associated with IFCN can be consulted in their website [6].

2.3.3 Data Annotation Methods

A large majority of the research that focuses on the characterization of unreliable content/accounts and the development of detection systems based on supervised learning techniques requires annotated data in some form. In the context of unreliable information, the data to be annotated can take a variety of formats, from the classification of websites into multiple unreliable and reliable labels [163, 147] to rumors extracted from Twitter based on newsworthy events [269].

When it comes to the annotation of unreliable information, there is usually a trade-off between cost and quality. While human-annotation methods often provide high-quality labels, these are usually assigned by domain experts. The number of these experts is often limited and thus obtaining large volumes of expert-annotated data is a difficult task. Therefore, a compromise is made in the quality of the annotation in order to obtain a larger amount of labeled data. Annotation schemes that follow this type of approach can be classified as distant or weak labeling. Both these types of annotation are discussed below.

2.3.3.1 Human Annotation

Human annotation is the most common type of annotation, but also the most costly (in terms of budget and time). Most studies in the literature that adopt this type of annotation resort to experts to perform the labeling. In the area of unreliable information, human annotation by experts has been conducted in different types of data: annotating websites and classified them into different reliable and unreliable categories [163, 147], assessing the truthfulness of social media claims [269], or assessing the truthfulness of online articles [76].

Expert annotation has the disadvantage of being expensive and time-consuming if the task requires constant annotations of the data. A weaker labeling option is to use crowd-sourcing

[3] https://www.washingtonpost.com/news/fact-checker/
[4] https://www.nytimes.com/spotlight/fact-checks
[5] https://fullfact.org/
[6] https://www.ifcncodeofprinciples.poynter.org/signatories

platforms. These annotation schemes are also being developed in the context of rumor threads on social media [268] with some evidence suggesting that they may be a reliable method to achieve trustworthiness labels in news media outlets [168]. However, there is a lack of this type of annotation data regarding unreliable articles and social media posts in the current literature. Crowd-sourcing platforms have also been used to annotate bot and human accounts [186].

2.3.3.2 Distant annotation

Distant-annotation or distant labeling is the most common type of annotation procedure for large quantities of data. In the context of unreliable information, a common strategy is to generalize more high-level annotations to low-level data. For example, the annotation of news articles [111, 126] or tweets [104, 31] based on the label associated with the news sources.

Figure 2.3 compiles the main data annotations used in current literature.

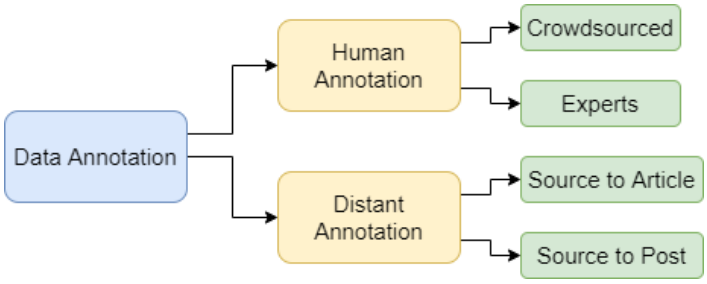


Figure 2.3: Diagram of the main data annotation methods presented in the current literature regarding unreliable information.

2.3.4 Datasets

Combining the sources and types of annotation, several datasets were created. These datasets are often used in tasks related to unreliable information detection. In this section, we present some examples of these datasets and their main characteristics.

Fake News Corpus [222] contains approximately 9.5M news articles extracted from the source available in OpenSources database [163] combined with the New York Times and Webhose English News Article articles datasets. The main reason for the additional sources was to balance the data, since OpenSources does not have a significant number of "reliable" news. The number of different domains included is 745 and the labels for each article correspond to its source's primary label in the OpenSources dataset.

Fake News Challenge Dataset Although this dataset does not contain annotations normally associated with unreliable information datasets, it was used for the Fake News Challenge Competition. The dataset consists of two different types of files. The first is composed of two fields, the id and the corpus (main body) of the news. The second file contains a news headline, a body id from the first file, and a label regarding the stance of the headline (i.e. if it is related to the body). It is important to note that each headline has multiple labels for different news corpora. The number of body ids for each headline is dynamic, ranging from 2 to 127. The main goal of this dataset was to develop systems capable of identifying if an article's body matches, contradicts, discusses, or is unrelated to the headline. This methodology can be extrapolated to determine whether a dubious news story is consistent with what well-established news media publish on the subject.

This dataset is an extension of the work of Ferreira et al [76] and their Emergent Dataset, which is described in the next subsection.

Emergent Dataset [76] consists of rumors (in the form of claims) extracted from multiple sources. These claims are then manually linked to multiple news articles where each news article is labeled by journalists regarding their stance on the original claim. Finally, after multiple articles are collected, a veracity label is assigned. For each headline, there are three possible labels: "for" (when the article is in accordance with the claim), "against" (when the article's statement opposes the claim), and "observing" (when no assessment of veracity is made by the article). The veracity label is initially set to "unverified". However, with the aggregation of articles regarding a claim this label is converted to "true", "false" or remains "unverified" if no evidence is found.

Kaggle Dataset This dataset [192] was originally published at Kaggle [7], a platform for machine learning and data scientists enthusiasts, where datasets are published and machine learning competitions are held. The dataset uses OpenSources and each post is assigned with its source's label. The dataset contains approximately 13 000 posts from 244 different sources. While this dataset follows the same methodology as the Fake News Corpus dataset, it adds a new label (bs) and a spam score provided by the bs-detector [227] application as well as some social feedback provided by the Webhose API [8].

Baly et. al Dataset This dataset was used in the work of Baly et. al [11] and was extracted according to the information available in Media Bias/Fact Check [147] which is a website similar to OpenSources where unreliable and reliable news sources are annotated. The labels provided on this website are manually annotated in two different categories: type

[7] `http://www.kaggle.com`
[8] https://webhose.io/

of content (for example "least bias" and "left bias") and factual score (for example "low", "mixed", and "high"). The dataset was created by crawling the website and retrieving approximately 1050 annotated websites.

BuzzFeed/BuzzFace Dataset On August 8, 2017, BuzzFeed published a study of partisan websites and Facebook pages and how they were proliferating after the 2016 United States presidential elections [216]. The dataset used for the analysis includes a collection of partisan Facebook pages and websites. These pages present elements of extremely biased opinions from both the left and right political sides. Although not all news stories included were false, they had extremely biased opinions and were aggressively promoted and disseminated on Facebook. In addition to the Facebook page name and website, this dataset also includes information about the registration date of the websites and whether they were linked to other unreliable websites (using Google Analytics and Google AdSense). In addition, engagement metrics regarding the different Facebook pages (likes, reactions, shares, and comments) are also included. Each website is classified as "right" or "left" according to its political leanings. Subsequently, an additional dataset, BuzzFace [197], was created containing complementary information such as replies to comments of the extracted original posts.

BuzzFeed-Webis Fake News Corpus 2016 This dataset presents a set of 1627 news stories from 9 different publishers retrieved in the 7 weekdays leading up to the 2016 US presidential election [177]. Of the 9 publishers, 3 are mainstream ("ABC", "CNN", "Politico") and 6 are hyperpartisan: 3 on the left ("addicting info", "occupy democrats", "the other 98") and 3 on the right ("eagle rising", "freedom daily", "right-wing news"). Each news article was evaluated by journalists regarding their veracity in 4 classes: "mixture of true and false", "mostly true", "mostly false", and "no factual content".

PHEME Dataset The PHEME dataset was first presented by Zubiaga et. al [269]. The authors collected rumors from Twitter based on 9 newsworthy events classified by journalists. Then, after capturing a large set of tweets for each event, a threshold was set and only tweets that had a significant number of retweets were considered for annotation. The annotation process was conducted by journalists and each tweet was annotated as "proven to be false", "confirmed as true" or "unverified".

LIAR Dataset The LIAR dataset presented and described in [251] is, to the best of our knowledge, the largest human-annotated dataset for false information analysis. The dataset is extracted from PolitiFact and includes 12.8K human-annotated claims. Each claim is labeled with one of the following six veracity degrees: pants fire, false, barely-false, half-true, mostly true, and true. The dataset was also sampled to check for coherence. The agreement

rate obtained using Cohen's Kappa was 0.82.

SemEval Dataset This dataset was created for the 2019 SemEval Task on Hyperpartisan News Detection [126]. This task consisted in detecting articles that follow a hyperpartisan argument. In other words, if it displays "blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person" [166]. The dataset is divided into two different types of annotations: by publisher (where the source of the news is annotated according to BuzzFeed journalists or the Media Bias/Fact Check website) and by article (annotated using multiple evaluations in a crowdsourcing fashion). The dataset has a total of 750 000 articles labeled according to their publisher and 645 articles labeled by crowdsourcing workers. The labels by publisher are the same as those assigned in Media Bias/Fact Check. The labels assigned to articles are simply "true" (if the article is hyper-partisan independently of the political side) and "false".

NBC Twitter Propaganda Accounts Dataset This dataset was collected by NBC News and includes 200k removed accounts that published and disseminated propaganda to influence the 2016 US presidential election [175]. The list of accounts was released by the United States Congress. Thus the majority of accounts were deleted, it was possible to restore a part of the data concerning tweets and account information. The data is divided into users and tweets with supplementary data extracted from the Twitter API (number of followers, tweets, and favorites for the accounts and number of retweets, hashtags, and mentions for the tweets). The main difference between this dataset and others is the inclusion of only one class of tweets/accounts.

MultiFC MultiFC [125] presents 34.918 claims extracted from 38 different fact-checking websites. Due to the multi-source extraction, the information provided varies with some articles including a label, a category, information about the person who made the claim, and a rationale for the label, while others are less comprehensive. Nonetheless, the dataset presents evidence pages to support the claims, which can contribute to a better understanding of the performance of veracity prediction models.

CREDBANK CREDBANK [153] is a tweet-based dataset containing over 60 million posts aggregated into 1049 real-world events. The dataset construction was divided into 5 steps. First, tweets were extracted using the Stream API and a topic analysis was performed using Latent Dirichlet Allocation (LDA). Next, the groups of tweets were sent to crowdsourcing for manual verification of the events. Then, credibility assessment of the different events was performed by crowdsourcing annotators using a likert scale from -2 (Certainly Inaccurate) to +2 (Certainly Accurate). This dataset differs from the previous

ones since credibility is annotated regarding events rather than specific posts or articles.

FakeNewsNet FakeNewsNet [209] is a repository/dataset of fake and real articles from two different fields: politics and gossip. The authors rely on a distant annotation approach using PolitiFact and GossipCop as fact-checking sources and extract the fake and real claims. Then, each claim is used as a query in search engines. In addition, each article is complemented with social context and spatiotemporal information to study the evolution of social feedback in real and fake articles on social networks.

FakevsSatire The "Fake News vs. Satire" dataset [87] includes political fake news articles and satire news articles that have been manually verified by annotators. In addition to the "satire" or "fake" label, each article was also assigned a topic such as "criticism towards a person or group", "conspiracy theories", or "sensational crimes". The dataset contains a total of 283 fake articles and 203 satirical ones.

NELA-GT-2018/2019/2020 NELA-GT [91] is a multi-labeled news dataset. The authors have released additional versions since its introduction in 2018. Data collection and annotation (distant-labeling) rely on Media Bias/Fact Check website. The most recent version (2020) includes 1.8M articles and additional information such as the tweets embedded in the article (as these are often the starting point for a news article).

FEVER FEVER [229] is a dataset containing approximately 185.5k claims generated from information extracted from the introductory section of several Wikipedia pages. Human annotators were responsible for generating the claims either by using real or false information. Then, each claim was labeled as "supported", "refuted" or "not enough info" by different annotators. This dataset differentiates from the previously mentioned since claims are created purposely towards the development of the dataset.

2.3.4.1 Dataset Comparison

With a large diversity of datasets tackling different tasks regarding the problem of detecting unreliable information, it is important to summarize the main characteristics of each like the content type, publication year, or number of entries. Table 2.3 presents an overall comparison of all datasets previously mentioned.

Table 2.3: Comparison between several state of the art datasets in the classification of unreliable information

Name	Content Type	Publication Year	Number of entries	Human Annotation	Labels
Fake News Corpus	News Articles	2020	9.5M	No	fake, satire, bias, hate conspiracy, state, junksci, clickbait credible , political proceed with caution
Fake News Challenge [70]	News Articles	2017	75 390	Yes	agree, disagree, discusses,unrelated
Emergent [76]	News Articles	2016	262 794	Yes	stance labels: for, against, observing veracity labels: unverified, true, false
Kaggle Dataset [192]	News Headlines	2016	12 999	No	fake, satire, bias, conspiracy, state, junksci, hate, bs
Baly et. al [12]	News Sources Websites	2018	1066	Yes	Factuality: Low, Mixed, High Bias: Extreme-Left, Left, Center-Left, Center, Center-Right, Right, Extreme-Right
BuzzFeed/BuzzFace [216]	Facebook Pages	2017	677	Yes	left, right
BuzzFeed-Webis [177]	News Articles	2016	1627	Yes	mixture of true and false, mostly true, mostly false, no factual content
PHEME [269]	Twitter Posts	2016	breaking news 9 conversational threads 330 tweets 4842	Yes	thread: rumor, non-rumor rumor: true, false, unverified tweets: support, denies, underspecified responses: agree, disagrees, comments
LIAR [251]	Claims	2017	12 800	Yes	pants-fire, false, barely false, half-true, mostly true, true
SemEval [126]	News Articles	2018	publisher 750 000 article 645	Partially	article-hyperpartisan : true,false publisher-hyperpartisan: true,false publisher-bias: left, left-center,least,right-center,right
NBC Twitter Propaganda [175]	Twitter Accounts	2017	tweets 267 336 accounts 513	No	None
MultiFC [125]	Claims	2019	34 918	Yes	Multi Domain dataset with multiple labels
CREDBANK [153]	Twitter Posts	2015	60M posts, 1049 events	Yes (in the events)	-2 (certainly inaccurate) to 2 (certainly accurate)
FakeNewsNet [209]	News Articles	2019	23 196	No	fake, real
FakevsSatire [87]	News Articles	2018	486	Yes	fake, satire
NELA-GT-2018/2019/2020 [91]	News Articles	2020	1.8M	No	**Reliability** unlabeled, reliable, unreliable, mixed **Factuality Score** unlabeled, very high, high, mixed, low, very-low
FEVER [229]	Claims	2018	185 500	Yes	"supported", "refuted", "not enough info"

2.4 Research Paths

Unreliable information presents a multidisciplinary problem with several lines of research currently being conducted. For example, some studies analyze the patterns and characteristics of unreliable posts in social media, whether over a large period of time [247] or on specific events [95, 218, 122]. Others converge this analysis towards bot accounts [201, 225, 138]. Another focus of research on unreliable information is the automatic detection of unreliable

content [6, 223, 171] or the detection of unreliable accounts [43, 243] .

The work in this manuscript seeks to make useful contributions to these specific lines of research and thus, the majority of this chapter will focus primarily on work on these topics. However, it is important to note that research on unreliable information is also being conducted in developing techniques for fact-checking claims [45, 205] or analyzing patterns of unreliable information diffusion in social networks based on epidemiological compartmental models [224]. A brief overview of research in these areas is discussed at the end of this chapter.

2.4.1 Exploratory Analysis of Unreliable Information

Most studies analyzing unreliable content and accounts on social media are conducted in relation to a certain event such as catastrophes [149, 228], terrorist attacks [95, 96, 218], or political events [31, 120, 16, 89]. Fewer studies are conducted with a broader perspective. An exception is the manuscript published by [247] which covers a large period from 2006 to 2017. Several findings are important to highlight. First, according to the authors, false information travels much faster through the network than real or credible news, beginning sometimes with a slow propagation, but once they become viral, their diffusion increases rapidly. Furthermore, unreliable posts tend to increase in important events such as elections. Concerning fact-checking or credible news diffusion, this study reaches similar conclusions as other related works (i.e. that false information propagates faster and in higher quantities than real news or fact-checking content). For example in [218], the authors claim that there is a misinformation to correction ratio of 44:1. This goes against previous findings [149] which support that there is a 1:1 false information to correction ratio. In a more recent study [203] the correction is 1:17 thus highlighting an absence of agreement on this subject.

Similar to content analysis, several studies examine social media accounts' accountability concerning the propagation of unreliable information on social media. In [212], the authors claim that accounts that trust unreliable content are registered earlier and have a higher following/followers ratio (i.e. accounts tend to follow more accounts than to have "followers"). Most studies also agree that social bots are involved in spreading unreliable information on social networks. It is estimated that the percentage of social bots in Twitter is around 15% [243]. Furthermore, in specific events (like elections or tragedies), they act like "super-spreaders" since several studies suggest that a large volume of tweets diffusing unreliable content is due to a small number of bot accounts [16, 202] and the majority of it happens at an early stage (i.e. a few moments after an unreliable news article is published for the first time). Social bots also have different strategies regarding the information they disseminate. A recent study analyzed the main strategies used by social bots in disseminating content on the awakening of an important event (Parkland shooting in Florida). The findings

suggest that 36% of the bots retweeted content that criticized the actors involved in the shooting (such as the police and the mainstream media). Other strategies used by social bots included fomenting doubt, sharing reliable information (showing that not all bots are malicious), spreading conspiracy theories, political organization, and commercial gain [127]. Nevertheless, it is important to emphasize that although social bots amplified discussion on social networks, it is the human-operated accounts that are largely responsible for the proliferation of bot-generated content [75, 127].

2.4.1.1 Case Studies

In the current literature, there are several examples of publications based on the analysis of social media activity in real-world events and the quality and credibility of the information shared in that medium at that time. Early work has focused on catastrophes like earthquakes [149], terrorist attacks [95] and nuclear disasters [228]. More recently, given the impact of unreliable information in the United States presidential election, politically motivated events have also been extensively analyzed. Furthermore, Covid-19 has also drawn the attention of researchers to study the reliability of information disseminated on social networks in the context of the global pandemic.

Due to the variation on the importance and impact of these events (whether due to their geographical context, absence of unreliable information being propagated, or overall relevance), some have been covered more extensively in the literature than others. We provide some examples of some of the most relevant ones in the context of unreliable information.

Boston Marathon Bombings The 2013 Boston Marathon was the target of a terrorist attack, in which two homemade bombs were detonated near the finish of the race [48]. This led to a rapid spread of information on social media, with several degrees of reliability. In [96] the authors conducted an analysis on 7.9 million tweets regarding the bombing. The main conclusions were that 20% of the tweets were true facts and 29% were false information (the rest were opinions or comments), it was possible to predict the "virality" of fake content based on the attributes of the users that disseminate it, and accounts created with the sole purpose of disseminating fake content often had names that resembled official accounts or names that elicited people's sympathy (using words like "pray" or "victim"). Another work looked at the main rumors spread on Twitter after the bombings occurred [218]. Here, the authors focused on 3 particular events and concluded that the volume of posts containing misinformation far exceeded the volume of posts correcting it.

Brexit Referendum The Brexit referendum was held in 2016 with the intention of allowing citizens to decide whether the United Kingdom would leave the European Union.

Because several claims during the campaign were dubious or lacked context [135], several studies have examined social media to understand what impact these claims had on the event. The authors in [16] discovered a large number of bot accounts that spread content mainly supporting the "Leave" vote, with some accounts retweeting campaign content for amplification of the campaign in social media, while others retweet active users on the network who also supported the "Leave" vote. Similar conclusions were described in [141] where the authors investigated accounts associated with state-sponsored Russian activity in the 2016 United States presidential election. According to the authors, these accounts significantly changed their behavior on the referendum day, disseminating content from other troll/bot accounts to amplify their impact (with content from the "Leave" campaign slightly more present in the tweets analyzed). These findings are reinforced in [113] where the authors concluded that hashtags associated with the "Leave" campaign are more than twice as prevalent as those associated with the "Remain" campaign and that less than 1% of the accounts sampled provide near a third of all tweets captured, suggesting extensive account automation.

2016 US Presidential Elections The 2016 United States presidential election would be the event that pioneered the expression "fake news" and raised awareness of the impact of unreliable information on social media. Consequently, several studies were conducted with respect to this particular event. The authors in [2] combined online surveys with information from fact-checking websites to determine the impact of unreliable information in social media and how it influenced the elections. Findings suggest that fake news articles in favor of Trump were shared three times more often than articles in favor of Clinton, and that the average American adult saw at least one false story during the month of the election. Another work [30] studied the influence of unreliable information and well know news outlets on Twitter during the election. The authors collected approximately 171 million tweets in the 5 months prior to the election and showed that bots spreading unreliable content are more active than those spreading other types of news (similar to what was found in [2]). In addition, the network diffusing false and extremely biased news is denser than the network that spreads center and left-leaning news. Other works regarding this event are presented in [131, 202] with very similar conclusions.

Coronavirus The coronavirus pandemic (Covid-19) has, once again, drawn attention to the implications of unreliable information dissemination (this time, in a health-related scenario). A multi-platform study was conducted in [47] where the authors analyzed the content regarding the pandemic on Twitter, Instagram, Reddit, Youtube, and Gab (a Twitter-like platform with fewer policies regarding shared content). With respect to unreliable information, the authors concluded that Gab accounts respond more to unreliable posts than to reliable ones, that Twitter presents a more neutral engagement with both sides

Table 2.4: Case studies regarding unreliable information on social media as respective works

Topic		Related Work(s)
Catastrophe	Fukushima Disaster	[228]
	Hurricane Sandy	[97]
Health	Coronavirus	[260, 74, 146]
	Others	[241]
Terrorism	Mumbai Blasts	[95]
	Boston Marathon	[218, 96]
Political	2016 US Election	[120, 90, 31, 2, 77]
	Brexit	[16, 113, 88, 89, 141]
	Other Elections	[46, 59, 133]

being very comparable, and that Youtube and Reddit show higher engagement with respect to reliable content. A more direct study of unreliable information is presented in [260]. Here, the authors investigated the information from tweets linking to several low credibility sources as well as two reliable sources: the New York Times, and CDC.gov [9]. The results suggest that low credibility content is more likely to be disseminated by bot accounts, although the percentage of retweets rate is higher for tweets containing links to the CDC.gov website and bot scores are higher on the "retweeter" accounts, suggesting that bots may also be responsible for disseminating reliable information.

Table 2.4 presents some of the most relevant case studies where unreliable information was studied in social networks and a sample of relevant and related publications.

2.4.2 Detection of Unreliable Content and Accounts

Regarding the detection of unreliable content in social media we can identify two main tasks. The first aims to predict if a social network post is false (or similar concept). More formally given a post with a list of predictors/features $\{X_1, X_2, ..X_n\}$ and a target variable Y, we aim at approximating the unknown function f such as $Y = f(X_1, X_2, ..X_n)$, with Y taking two possible values/labels (e.g. False/True, unreliable/reliable or True/Fake News). In some applications, the target variable Y can have more than two values (fake/reliable/satire) making it a multi-label classification task.

The second task concerns the identification of bot accounts that play an important role in the dissemination of unreliable content on social media. It is usually approached as a classification task (i.e. to label an account as being a bot or human) [57] although there are also studies that approach the problem as a multi-label classification task since they consider an intermediate type of account (cyborg), which is a mainly automated account with rare human intervention [43].

[9] Center for Disease Control and Prevention

Although content detection and bot detection are the main tasks addressed in this work and are considered to be the ones that contribute more directly towards the mitigation of unreliable information online, stance detection is also a well-studied area with a large set of works in the domain of unreliable information detection in social media. We briefly discuss this task in the following paragraphs.

In previous research, stance detection has been defined as follows: given a fragment of text the task aims at predicting if it agrees, disagrees, or is unrelated to a specific target topic/claim. However, in the context of unreliable content detection, stance detection has been adopted as a primary step to detect the veracity of a news piece.

The connection between stance detection and unreliable information was reinforced by the previously mentioned Fake News Challenge which promoted the identification of unreliable information through this task. This task can be generalized out of the scope of the challenge towards the effectiveness and speed of fact-checking since a claim can be inserted as input and quickly retrieve articles that supported and refute it, allowing quicker reasoning and assessment on its veracity [103]. Several approaches were presented in the context of the competition. The authors in [156] present a set of experiments using a conditioned bidirectional LSTM (Long Short Term Memory) and the baseline model (Gradient Boost Classifier provided by the authors of the challenge) with an additional variation of features. Other works with similar approaches have been proposed [172, 206]. However, the results achieved do not vary significantly.

Stance Detection is also used to identify how a piece of unreliable information spreads in social networks. More specifically, the task has contributed toward the detection of rumors by using the wisdom of the crowd to assess the veracity of a claim. The works of [63, 144] are a few examples of the application of stance detection towards the identification of rumors in social networks. A more comprehensive review is presented in [103].

2.4.2.1 Input Features

When applying machine learning techniques, it is necessary to analyze and select important features that are able to distinguish between different class labels. Although, different social networks have different characteristics, in the current literature there are three main data tiers from where features can be extracted and which are common to the different platforms: posts/publications, accounts, and network information. The majority of these data are easily accessible through the API of the respective social network (if available). Account information such as the handler/username, creation date, and profile picture as well as post information (text, date, number of likes/favorites, number of shares) are usually extracted instantly through the API and require a single call. However, when we look at components such as the account timeline and connections, these are dynamic in size as there are accounts

(a) Account profile *(b)* Post information *(c)* Network Information

Figure 2.4: Several components from social media from where input features can be extracted.

with millions of connections and posts and accounts with only a few. Consequently, the extraction time increases with the amount of information in the accounts and the degree of completeness of the information needed. For the accounts timeline, the extraction time is linear with the number of posts of the accounts, being only subject to the restrictions of the API. With respect to network information (i.e. the number of followers and friends), this can be linear for direct connections or increase exponentially depending on how complete the information should be (i.e. only friends/followers, friends of friends, ...).

Figure 2.4 visually illustrates the different social network components and the common information available to derive features from. In the following subsections, we describe the features that can be extracted from each component. It is important to highlight that both unreliable accounts/bots and unreliable content tasks share most of the described features. Therefore, we focus on dividing them by social network components instead of tasks to avoid repetition.

Post-based features Post-based features are mainly extracted from the text of the publication, media, and social feedback (likes, comments, replies...). With respect to the text, text statistics are frequently considered useful features for detecting unreliable content. These include text length, number of words/characters, and percentage of uppercase letters. Punctuation is also commonly used whether in boolean (e.g. the presence of an exclamation or question mark) or numeric form (e.g. total number of question marks or ratio of exclamation marks). In addition, specific elements of social-network are also often considered, namely the number of hashtags, mentions to other accounts, or external links (URLs).

Natural Language Processing (NLP) techniques are also used to extract useful information from the post's text. Parts of speech (POS) tags (such as the number of nouns and the presence of pronouns), sentiment (e.g. the number of positive and negative words), and entities (such as people, organizations, and locations) are some examples of the features conveyed by several studies in the literature [29, 245, 130, 149, 104]. In addition, some studies [101, 104, 245, 120] also use a Bag of Words approach or Word Embedding models

to create a large set of features based on the text of the post. Additional features can be extracted recurring to other tools. The psychological meaning of words can be analyzed using the LIWC tool [226] and can be used as a feature based on the psycho-linguistic characteristics carried over in unreliable and reliable posts. Additionally, readability scores, link credibility via WOT score [10], and Alexa rank [11] can also be used as input features [29].

Finally, the integration of media in social media posts allows a large number of additional features. Although there is a wide study area on identifying fake/manipulated images on social networks, works such as [122] combine text and image features for the identification of false and real news in tweets. The authors proposed visual features such as clarity score, coherence score, and diversity and clustering score and show that the combination of these features with more traditional based ones yields a better overall performance on the detection of false and real news tweets. In another work [252] text and image features (extracted from the pre-trained VGG-19 neural network) are combined to detect fake news and discriminate between different events. A similar work also relies on a multimodal approach, using VGG-19 and word embeddings to extract image and text features (respectively) for rumor detection in social networks [119]. However, these features can only be applied to posts that contain this type of media.

Account-based features The features extracted solely based on account information include the number of followers and friends, verification status, account age, and number of posts [104, 149, 130]. In addition, the absence/presence of biography, profile picture, and banner are also used [29]. Depending on the social network where the data is extracted, specific features can be retrieved. For example, some studies use the gender of the account and the type of username which is available on the Weibo social network [259, 254, 264].

Features extracted from the posting history of each account can also be integrated into this category. These usually consist of post-based features performed on all the accounts' publications followed by some kind of aggregation function. For example, the number of hashtags is extracted for each post and the average per post is sometimes considered as an input feature. Similar procedures are applied to sentiment, number of entities, and hashtags with functions such as average, sum, count, and standard deviation frequently considered for the aggregation process. The history of each account's publication also gives way for additional features to be extracted. For example the time between two consecutive posts or repeated publications can be an important feature for the bot detection task. In addition, word-embedding models can also be used to extract meaningful features from the accounts' publication history.

Another interesting approach to feature extraction which is still rare in the literature is the

[10] https://www.mywot.com/

[11] https://blog.alexa.com/marketing-research/alexa-rank/

concept of extracting features with respect to the account evolution process. The work in [263] proposes these features by monitoring a set of accounts on a daily basis. Consequently, additional indicators can be extracted such as screen name changes, the increase or decrease in the number of friends and followers, and the number of tweets deleted over time.

Network Propagation Features Groups of features that are less commonly used in detection tasks include propagation-based and link-based features. For example, features based on the analysis of cascade of retweets such as depth, maximum sub-tree, and maximum node are proposed in [38]. In addition, the work of [254] also presents additional propagation features based on the reposts of the original post. In this work, features such as the average sentiment, doubt, or surprise of the reposts, as well as the interval between the original message and the reposts are considered.

2.4.2.2 Model Types

The different tasks regarding unreliable content and accounts are commonly presented as text mining classification tasks. Therefore, the models and metrics used for them are similar to those used for other text classification tasks (e.g. sentiment analysis, document classification).

As mentioned earlier, both tasks often consider a binary classification scenario with an unreliable label (such as "false" or "bot"), and a reliable one (such as "true" and "human"). Thus, the machine learning models used are similar. Several studies have used logistic regression, Decision Trees, Support Vector Machines, Naive Bayes (most frequently on bot detection task), and K-Nearest Neighbors. We describe each model (that later will be used in Chapter 4.2 and 5) briefly and simply in the next paragraphs:

Decision Tree. Decision Tree [183] is a tree-based model where each node has an associated feature and condition. Commonly, information gained is used to select each condition for each node (although several split criteria can be applied). Then, for each new prediction, the tree is navigated starting from the root node until one of the leaves (which contain the possible labels).

Support Vector Machines. In classification tasks, SVMs [51] attempt to separate examples from both classes using a hyperplane and maximizing the margin distance between classes. SVM models can be Linear or Non-Linear. In Linear SVM, the assumption made is that both classes can be separated by a linear space. In Non-Linear SVM, a kernel function is used to create a new hyperplane so data can be separated in a linear fashion.

K-Nearest Neighbors. KNN [4] is a model that uses the neighborhood to classify a new entry. More specifically, the most similar k data points are used. Finally, majority voting over the k-neighbors' classes is performed to determine the new node class.

Naive Bayes. Naive Bayes [102] is a generative model that relies on the Bayes rule to determine the label of an entry given the input features. Different Naive Bayes models exist based on the assumptions of the distribution of the data. It is called 'naive' because it assumes total independence of the variables.

In Equation 2.1 the Bayes rule is presented. Y is the output label while $(x_1, x_2, \cdot x_n)$ is the set of input features.

$$P(Y|(x_1, x_2, \cdots x_n) = \frac{P(x_1, x_2, \cdots x_n|Y).P(Y)}{P(x_1, x_2, \cdots x_n)} \tag{2.1}$$

Thus, the likelihood of a label given the input features $(x_1, x_2, \cdots x_n)$ is determined by the likelihood of the features given the label, the likelihood of the label, and the likelihood of the features.

The performance of traditional machine learning models is often improved with the use of ensemble models. These consist of a combination of multiple models where each model's output contributes towards the final prediction. A simple ensemble approach can be to use majority voting to make the final prediction based on a set of different machine learning models. We briefly describe some of the most commonly used ensemble approaches in the previously mentioned tasks.

Random Forests. Random forests are constructed using an ensemble of decision trees. Each decision tree is trained using a random subset of the training data as well as a subset of the input features. The final prediction is made based on the class that has the highest voting (whether it is by count or average probability).

Adaptive Boosting. Adaptive Boosting [78] consists of an ensemble of weak models arranged in a sequential fashion where each new model learns taking into account the errors of the previous models. Usually, the models used are decision trees with only one node and two leaves (also called decision stumps).

Gradient Boost. Gradient Boost [79] for classification tasks adopts an approach similar to Adaptive Boosting. However, decision trees are generally used (instead of decision stumps). In addition, decision trees are used to predict the residual instead of the label itself. The

residual predictions of the different trees are then used in combination with a learning rate to make the final prediction of the label.

The previously described models are the ones that are commonly used for both the tasks of bot detection and unreliable content detection. However, in recent years, more complex models have been adopted to tackle the different tasks in unreliable information. For example, [230] uses Long-Short Term Memory (LSTM), Convolutional Neural Networks (CNN), and BERT for stance detection in Twitter. Similarly, [214] proposes a combination of Nested LSTMs as well as Densely Connected Bidirectional LSTM in the same task. Finally, in the previously mentioned Fake News Challenge, several studies proposed different versions of LSTMs to solve the problem [172, 206, 184].

In the fake news detection task, the use of deep learning models has been applied to the classification of articles from fake and real news websites. Kalyar et al. [123] used several word-embeddings and BERT encoders with machine learning and deep learning classifiers in news articles propagated during the 2016 United States presidential election. In addition, BERT was also used to detect fake news spreaders [161] and for the classification of false and real claims [139].

Nevertheless, appropriate evaluation metrics are needed to measure the performance of the models. The different evaluation metrics used in the tasks described in this chapter are discussed in the following section.

2.4.2.3 Evaluation Metrics

In traditional machine learning, a detection model/system, given an input (p.e. an account or post), will produce a prediction based on the labels provided in training. In binary classification tasks, a total of four different scenarios can occur depending on the true label and the predicted label. A correctly predicted instance is considered a True Positive (TP) or True Negative (TN), depending on which class we consider (for example, the positive class can be associated with the "reliable"/"human" class while the negative may be associated with the "unreliable"/"bot" class). On the other hand, an incorrect prediction can be defined as false positive (FP) in the cases where a negative instance is predicted as positive and false negative (FN) in the remaining scenario.

The results of applying a machine learning model to a set of test cases are usually presented in a confusion matrix like the one shown in Table 2.5.

The number of true positives, false positives, true negatives, and false negatives are normally computed. Based on these numbers, several evaluation metrics used in classification tasks such as those addressed in this work can be calculated. Two of the most commonly used evaluation metrics are Accuracy and F-Score. Accuracy measures the rate of correctly

Table 2.5: Example of a confusion matrix for binary classification.

		Prediction	
		Positive	Negative
Ground Truth	Positive	TP	FN
	Negative	FP	TN

predicted instances in both classes as shown in Equation 2.2.

$$\text{Accuracy} = \frac{TP + TN}{FP + TP + FN + TN} \tag{2.2}$$

F-Score is a more complex metric that uses Precision and Recall (or True Positive Rate). Both are calculated with respect to a single class, as Precision attempts to measure the performance of the model with respect to positive instances while Recall measures the ability of the model to correctly classify all positive instances. Precision and Recall are expressed in Equation 2.3 and 2.4, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.4}$$

Therefore, the F-score is the combination of Precision and Recall and a β coefficient that weights the importance of each component. The equation is shown in Equation 2.5 A $\beta > 1$ increases the importance of Recall while $\beta < 1$ increases the weighting of Precision. In the revised literature, the majority of papers use $\beta = 1$ which corresponds to a balanced F-score and is usually referred to as F1-score or F1-measure.

$$\text{F}_\beta = \frac{(1 + \beta) \times \text{Precision} \times \text{Recall}}{\beta^2 \times Precision + Recall} \tag{2.5}$$

It is important to emphasize that the F1-score is computed regarding a particular class. However, the majority of works in the literature focuses on the performance of the system with respect to both classes (reliable and unreliable). In addition, Accuracy presents some limitations if the classes in the test data are imbalanced. In other words, if the percentage of one class in the test set corresponds only to 10% of the data, a model that always predicts the other class will have an accuracy of 90% which can be misleading. Similarly, F1-score can also be misleading if these values are only presented regarding a single class and both classes are relevant to the problem.

To avoid misleading results, several solutions are possible. For example, information regarding F1-score for each class can be presented. However, combined metrics can also be used such as F1-score macro, micro, weighted average variations, and balanced accuracy. F1-score macro average can be used to evaluate the overall performance of the model disregarding classes size, while micro and weighted variations consider the possible variance in the number of entries in each class. Studies such as [125] use the macro and micro F1-scores for multiple domain classification, while in works such as [104] only the unreliable class is considered. Balanced accuracy works similarly, where the correctly predicted instances are weighted by the size of each class.

Finally, the Area Under the Curve (AUC) metric is also frequently used. This metric is based on the area under the ROC (receiver operating characteristic) curve which is based on the True Positive Rate (Recall) and the False Positive Rate (which is expressed in Equation 2.6).

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{2.6}$$

The ROC curve is based on probabilistic predictions by varying the decision threshold and the corresponding trade-off between True Positive Rate and False Positive Rate. The AUC, as the name suggests, is simply the area under the ROC curve. However, in the revised literature, the AUC is often achieved not only by the probabilistic predictions but by a single binary predictor. Consequently, Equation 2.7 can be applied.

$$\text{AUC} = \frac{1}{2} \times 1 + TPR - FPR \tag{2.7}$$

Models that have performed well in more traditional text mining tasks have been adopted in the context of the tasks discussed in this work. For example, the studies of [38, 130, 101] use Decision Trees and achieve an F-score between 0.83 and 0.86. On the other hand, [264, 254, 259] and again [130], use Support Vector Machines for the unreliable content detection task, accomplishing F1-scores between 0.74 to 0.90. Other approaches include the use of ensemble models (0.9 f1-score) [104] and Convolutional Neural Networks (0.95 accuracy) [245]. A more unusual approach is the harmonic boolean label crowdsourcing presented in [223] which relies on social feedback from users to predict whether a post is a hoax or not. Although the authors describe excellent results (99% accuracy), the presented model relies on crowdsourcing the opinion of users based on past behavior. Therefore, it does not seem possible to apply this model in the absence of social feedback, making it unsuitable in an early detection scenario.

The problem of detecting unreliable users/bots is also generally addressed as a classification task, where similar classification algorithms are tested and evaluated. In several studies,

Random Forests achieve a good performance in distinguishing human and bot accounts with F1-scores ranging from 0.91 to 0.96 [10, 85]. Furthermore, the same model proves to be efficient in a three-label classification scenario (human, bot, and cyborg) achieving an AUC score of 0.95 [43]. Naive Bayes is also an often-used model accomplishing similar results [10, 67].

2.4.3 Network Based Solutions

In this section we briefly describe how network-based approaches help analyze and detect unreliable accounts and content. Because social networks are themselves networks, they share several characteristics and similarities with more traditional networks. Therefore, several studies transfer solutions applied in other domains to the unreliable information problem. More specifically, when analyzing unreliable content from a network propagation perspective, the problem can be compared to the spread of an infectious disease where a node (account) can be "infected" with a certain probability [124]. This probability may vary according to several factors. First, not all accounts believe in unreliable content thus it is important to divide them in three classes: the "persuaders" whose goal is to spread and support unreliable content, the "gullible users" who are easily influenced by unreliable content, and "the clarifiers" who are immune to unreliable publications and can confront infected users with fact-checking content [207]. Homophily and social influence theories contribute towards the importance of the friends' network in the "contamination" of a gullible user. Accordingly, the probability of a user believing unreliable information can be computed depending on the beliefs of the friends (i.e. a user who has friends that believe in unreliable content has a higher probability of being infected) [255]. Based on this approach, several models and user roles have already been proposed. Tambuscio et al. [225] develop a model for rumor spreading based on similar user roles (Believer, Fact-checker, Susceptible) and three probabilistic phenomena: spread (when the user spreads the rumor), verify (when the user fact-checks the rumor), forget (the user forgets the news). Another study [138] considers competing information spreading simultaneous in the network (i.e. the simultaneous spreading of unreliable and reliable content). Furthermore, time is an important factor in this model as the probability of a user reading an unreliable post from its close connections decreases over time.

Several important findings emerge from these studies. First, fact-checking activity on the network does not have to be very high to stop the spread of unreliable content, and even if it is removed from the network, fact-checking continues among users who believe it [225]. Second, the percentage of users protected from unreliable information increases when the propagation time constraints are more relaxed, and it is lower when the time constraints on dissemination of information are more restricted (i.e. when it is urgent to disseminate content, more users are infected) [138]. These results support and help to explain the results

in other previously mentioned studies. Namely that, in the occurrence of an event, the diffusion of unreliable information tends to occur in greater quantities [247] and that human accounts are mainly responsible for its spread [75, 127].

Another use of network structures in the context of unreliable information detection is the use of knowledge graphs for automatic fact-checking. These graphs usually use triples (subject, predicate, object) to represent facts. In a knowledge graph the nodes represent subjects and objects while the predicates are the edges.

As stated in [266], there is an absence of a full system for automating fact-checking based on knowledge graphs. However, two main tasks are identified on current research that can help towards that goal: fact-extraction and fact-comparison. Fact-extraction relies on the extraction and process to adapt raw and trustworthy information from the web to a knowledge graph. The second, fact-comparison, focuses on the verification of claims using knowledge graphs.

Some studies deal with the development of knowledge graphs. The authors in [45] treat fact-checking as a network problem. By using Wikipedia infoboxes to extract facts in a structured way, the authors propose an automatic fact-checking system based on the path length and the specificity of the terms of the claim in the Wikipedia Knowledge Graph. The evaluation is performed using assertions (both true and false) from the entertainment history and geography domains (for example "x was married to y", "d directed f" and "c is the capital of r") and an independent corpus with novel statements labeled by human annotators. The results of the first evaluation showed that true statements have higher truth values than false ones. In the second evaluation, the authors showed that the values from human annotators and the ones predicted by the system were correlated. Another work by the same authors [205] uses an unsupervised approach to the problem. The Knowledge Stream methodology adapts the Knowledge Graph to a flow network because multiple paths may provide more context than a single path and reusing edges and limiting the paths where they can participate may limit the path search space. This technique, when evaluated in multiple datasets, achieves similar results to the state of the art. However, in various cases, it provides additional evidence to support the fact-checking of claims. A different work targeting specifically unreliable content is presented in [167] where the authors build three knowledge graphs based on real and false news articles. Then, these are converted to an embedding model to be able to predict unreliable articles. Additional studies [244, 80] have complemented this fact-checking approach with explainability to provide a clearer and more easily understood classification.

2.5 Discussion

The ongoing work developed in the research community has targeted the multi-domain problem that is identification of unreliable information online. More specifically, a large effort has been made towards the identification of unreliable social media posts as well as the identification of bots and human accounts. On the other hand, from an applicational point of view, there are still limitations transitioning from experimental setups to complete solutions. With respect to the two main problems addressed in this work, we argue that the majority of studies in the state of the art concerning the detection of unreliable accounts is focused on the identification of bot accounts. Nevertheless, some studies work towards a different characterization of social network accounts regarding their reliability. For example, the work in [212] refers to users that are likely to trust false and real news and analyses characteristics to differentiate these accounts. Another work [99] uses the terminology "fake Twitter accounts" and "fake users" but are essentially bots since "are generated intentionally, and often automatically/semi-automatically, by cyber-opportunists (or cyber-criminals)". A similar concept that is often used is "spammer". However, the definition of spammer is often associated with the distribution of unwanted content such as advertisements, viruses, and pornography [20, 114]. Regarding accounts that disseminate unreliable content, current studies are very driven towards the analysis of the characteristics of these accounts. Examples include the work in [212] where a characterization of users' profiles is conducted and the work [207] where an evaluation of user-based features is conducted for the task of fake news detection. However, there seems to be a lack of studies regarding the automatic identification of users that disseminate unreliable content. A notable exception was the task proposed at CLEF 2020 for Profiling Fake News Spreaders on Twitter. In this task, the main goal was to "discriminate authors that have shared some fake news in the past from those that, to the best of our knowledge, have never done it" [185]. Nevertheless, only 100 tweets for each user were provided as training data. Several works emerged from this competition, with the best works achieving an accuracy score of approximately 77.5% [33, 173]. In addition, looking at a more realistic and application-based scenario, current solutions are limited to bot detection systems such as Botometer [243] and BotSentinel [12].

Regarding the detection of unreliable posts in social networks, the current literature mainly focuses on the development of supervised models to detect unreliable posts in specific real-world events, which often refer to data extracted in a small time interval. Consequently, longitudinal evaluation of the performance of the models over time with possible changes in the topics discussed is often not considered. Examples of these works include [104] where the authors develop supervised models to detect fake news, but disregard an evaluation over time or a temporal split between training and validation data. Similarly, in [6] which focuses on the detection of misinformation regarding a specific event (Hurricane Sandy).

[12] `https://botsentinel.com/`

Time can have a crucial impact on the performance of supervised models. More specifically, changes on the distribution of features or the concept/meaning of what is been classified can change over time, affecting the performance of the model. A common area where this phenomenon occurs is in recommender systems, where user interests change over time. These occurrences are commonly known as concept drifts and occur when the relationship between input variables and labels changes [82]. Two types of concept drift can occur.

- Real Concept Drift - refers to changes in the conditional distribution of the target label, while the input features distribution may or may not change.
- Virtual Concept Drift - occurs when the distribution of the input features change without affecting the distribution of the target label.

In the case of unreliable content detection over time, the latter definition applies because variation in topics may affect the distribution of input features. Real concept drift does not apply in this task because the concepts of what is reliable and unreliable do not change over time.

A work that addresses the concept of longevity of fake news detection models over time is presented in [112], where the authors explore the impact of phenomenona such as concept drift. However, the authors focus on articles from unreliable and reliable websites rather than posts on social networks.

The development of a system that can detect unreliable content over time must be aware of the drifts that may occur. To the best of our knowledge, little to nothing in the current literature addresses evaluating the performance of unreliable detection models in social networks over time, which could be useful when transitioning from a more experimental design to a fully developed and deployed machine learning-based detection system. Likewise, when addressing the unreliable account detection problem, few studies exist outside of the bot detection task. In addition, there seems to be a lack of studies that address the variation in the volume of content in social network accounts and how this may affect feature extraction and model performance. The closest approach to this problem is the work in [22] where the authors extract features and develop supervised models for bot detection using different data tiers. However, this paper concerns the bot detection task and the tiers used are limited to the inclusion/exclusion of API data collection tools (e.g., account information + 1 tweet, account information + timeline, account information +timeline + friends timeline) rather than different volumes of tweets.

Therefore, the work presented in this thesis attempts to bridge some aspects of current experimental approaches with real-world scenarios in both unreliable accounts and content detection. In particular, we address the problem of limited content in account timelines

for the classification of reliable and unreliable accounts and analyze the performance of supervised models in detecting unreliable content in social networks over time.

Chapter 3

Data Extraction and Preliminary Exploratory Analysis

In this chapter, we present the methodology used for data extraction in the experiments conducted in Chapter 4 and 5. As previously mentioned, we focus on identifying accounts and posts that contain unreliable content. This definition is less restrictive than the common definitions of "fake news" or misinformation, as it includes other types of content that go beyond the definition of "factual" news and thus can falsely influence users' opinions.

3.1 Data Extraction

In Chapter 2, we mentioned the different concepts that have been addressed in the relevant literature, both from an analysis point of view as well as on the development of detection models to identify content that meet such a definition. Some of the concepts discussed included were "fake news", extremely biased and hate content, junk science, conspiracy theories, and rumors. Although some of these concepts have motivated their own research topics with a large number of related publications (e.g. fake news and clickbait), in this work we focus on a more abstract concept that we define as unreliable information. Our motivation behind such decision, is twofold. First, the use of a term such as "fake news" or "misinformation" would increase the number of definitions associated with these terms as there is no consensus in the literature on a formal definition for each. Second, because we analyze accounts, with a less restricted definition of what is unreliable content, accounts that disseminate multiple types of unreliable information can be included in this work, allowing us to pursue a broader solution to the problem.

Regarding data sources, from the different social media platforms, we opted to conduct our work in Twitter. The main reasons for this decision are the following: 1) Twitter is one

of the main platforms where the spread of unreliable content has been reported, severely affecting users' perceptions of important events (such as elections) and general trust in science (e.g. vaccination against Covid-19). 2) Although Facebook has a larger user base, data accessibility was affected by leaks during the Cambridge Analytica scandal [35]. Since then, Facebook has restricted access to its API [68] and only recently (early 2021) created an API for researchers [69]. Twitter, on the other hand, has always maintained an open API for research purposes. Therefore, data extraction and dissemination are easier to accomplish.

The data annotation scheme chosen for this work had to take into account the nature of research conducted. In particular, as one of the tasks is do develop unreliable content detection models and assessing their longevity, large amounts of data spanning a long period of time were required to ensure better robustness of the results presented. This factor restricts the annotation process options due to the time-consuming task of manually annotating a large number of posts on a daily basis. Accordingly, the better choice for this task was distant labeling similar to the one conducted in previous studies (e.g. [112, 104, 13, 174]).

To guarantee the effectiveness of our distant labeling approach, we rely on OpenSources and MediaBias/FactCheck. Although we briefly mentioned these in Chapter 2, we detail further the characteristics of both sources in the next paragraphs.

OpenSources is a database of online information sources. Although the website (`http://opensources.co`) has been shut down, the database is still used by the scientific community as ground truth for several studies (as mentioned in Chapter 2) and has been the basis for several articles' datasets using a distant labeling approach. The database is currently available through the Github repository [1]. In OpenSources, each source can be annotated to a maximum of 3 (out of 12) categories in a ranked fashion (i.e. from the most predominant type of content to the least). However, in this work, from all the 12 different categories, we only select 5 since: 1) we are interested in using this database only for unreliable content (thus we exclude categories such as "political" and "reliable") and 2) other categories are less representative of the problem (such as "gossip"). In addition, we simplify our classification by using only the predominant tag on each website, adjusting in some specific situations where the second label provides an important characterization of the content. A common example is when the first label is "political" and the second is "fake" or "unreliable". In these cases, the second label is considered.

On the other hand, MediaBias FactCheck has two different types of classification for each source. The first is based on the type of content that each source publishes. The labels include 6 different bias values ranging from the extreme right to the extreme left (left,left-center,least,right-center,right). In addition, sources that fall outside of this spectrum, are given labels such as "Pro-Science", "Conspiracy/Pseudo-Science" or "Satire". The second

[1] `https://github.com/OpenSourcesGroup/opensources`

type of classification relates to the factuality of each source. The classification includes 6 labels, ranging from "Highly Factual" (when the source only presents factual content) to "Very Low" (when the source never uses credible sources and is not trustworthy).

We decided to use both databases because there is a limited number of reliable sources available in OpenSources (compared to the unreliable categories) and they are almost exclusively targeted at the political domain.

Table 3.1: Distribution of websites per unreliable class in OpenSources.

Classification	**Number of websites**
Bias	133
Hate	29
JunkSci	32
Fake	237
Clickbait	32
Unreliable	56
Total	**522**

Since our definition of unreliable includes several concepts such as extreme bias and false information, we can adapt some of the labels from OpenSources and aggregate them towards our definition of unreliable content. Therefore, in this work, we rely on several definitions provided by OpenSources to determine what unreliable content is. Thus, we define that a post/tweet disseminates unreliable content if it contains a hyperlink (URL) to a website whose content falls under one or more of the following definitions provided by OpenSources:

- **fake**: the content provided is fabricated information or distorts actual news with the aim of deceiving users
- **clickbait**: the content provided has an eye-catching title or headline with the sole intention of enticing users on social media to click on the associated URL
- **bias**: the content provided is extremely biased and aggressively favors the opinion of one side and/or demeans and insults opposite opinions.
- **junksci**: the content provided refers to scientific theories that are false or whose veracity is unclear (also known as junk science)
- **hate**: the news content provided promotes racism, homophobia, or other forms of discrimination.
- **unreliable**[2]: the content provided is unclear and requires further investigation to determine its veracity.

[2]In this work we use the term unreliable to classify the content provided by all categories. This is the only exception since, as we choose to leave the names of the OpenSources categories unchanged.

The number of different websites per unreliable label is presented in Table 3.1.

On the other hand, it is also important to define what is reliable content. The concept is similar. However, due to the limitations of the OpenSources platform in terms of reliable sources, we choose to use the MediaBias FactCheck (MBFC) [147] database and determine the following labels as reliable:

- Pro-Science - Sources that consist of legitimate scientific content and are based on credible scientific methods and sources.
- Left-Center Bias - Sources with minor democratic bias that are generally trustworthy for information.
- Least Bias - The most credible media sources with minimal bias and highly factual reporting.
- Right-Center Bias - Same as left-center bias but skewed toward more conservative causes.

The number of different websites per reliable label is presented in Table 3.2.

Table 3.2: Distribution of websites per reliable class in MBFC.

Classification	**Number of websites**
Pro-Science	129
Left-Center Bias	429
Least Bias	342
Right-Center Bias	207
Total	**1107**

We can see a large difference in the number of reliable and unreliable websites. However, this does not necessarily translate to a larger volume of tweets as less relevant or well-known information sources are less likely to be shared on Twitter. The vast majority of sources included have factuality scores ranging from "Mostly Factual" to "Very Highly Factual". Although we could restrain our sources solely to the ones with "Very Highly Factual" label, this would lead to a high large disparity between the number of reliable sources in the different categories, which would affect the diversification of the content considered and exclude some of the mainstream news sources (for example the New York Times, BBC and The Economist were labeled "Highly Factual" at the time of the analysis). Therefore, we relax our criteria slightly to ensure a more balanced number of sources between classes and consequently diversify the type of content extracted.

Each website's URL from the selected categories was used as a keyword for the Search API on Twitter. For each query/URL, a collection of 100 tweets is extracted daily. This allows

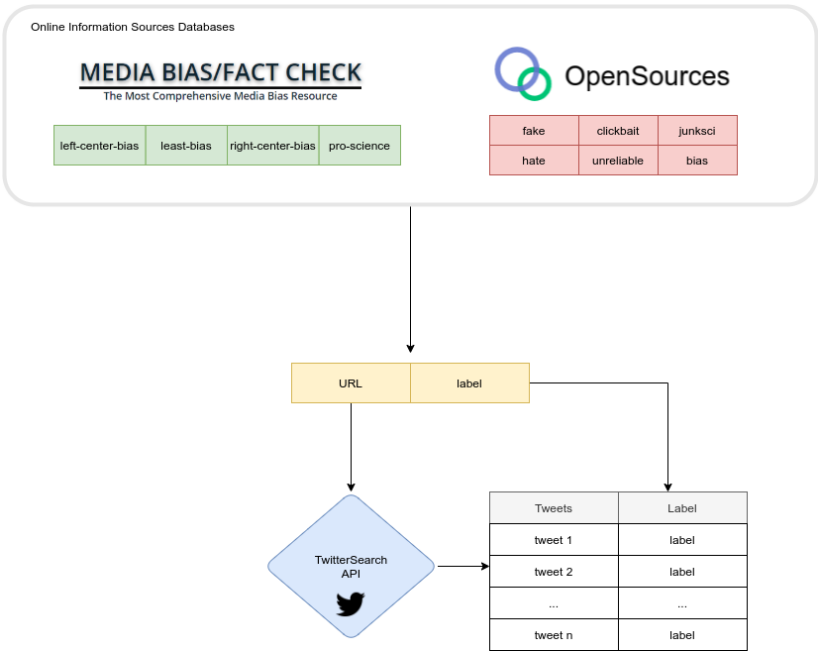


Figure 3.1: Data extraction workflow. For each pair (URL,label) present in MediaBias FactCheck and OpenSources databases, the URL is passed down to the Twitter Search API and collection of tweets is extracted that contain the queried link. Finally, the label is associated with each tweet.

extraction of tweets that contain the queried domain (or a sub-domain) in the text of the tweet. Consequently, by propagating links that connect to reliable/unreliable websites, the assumption step on this distant labeling approach is that the tweet itself is also propagating reliable/unreliable information. However, in some cases, this might not be necessarily true (for example when a tweet points out the misinformation of the disseminated link). Nevertheless, due to the large amount of data extracted and the success of previous studies using this method, these cases appear to be negligible, so this method also appear to be robust for large quantities of data (further analysis is also provided in Section 4.1 regarding the covered accounts).

In addition to the text of the tweet, social feedback such as the number of favorites and retweets as well as the account information (e.g. screen name, number of followers, verification status) were also extracted. Some mechanisms were also implemented to ensure that unique tweets were retrieved. In the case of repeated tweets, only the fields that suffered any change are updated (for example the number of retweets/favorites and the number of followers/followers of the account). Finally, each tweet was labeled with the source/URL label as well as a binary label (reliable or unreliable). For better understanding, Figure 3.1 presents a visual representation of the data extraction workflow. In addition, Figure 3.2 presents some example tweets extracted using our methodology.

Figure 3.2: Examples of tweets retrieved using the data extraction methodology. The links used in the query are present on the text of the tweet.

3.1.1 Account Annotation

We described the annotation process for each tweet in the previous section where the annotation of the website source is applied to the tweet.

To adapt our current data extraction workflow to annotate unreliable and reliable accounts, we assigned the numeric label -1 to the unreliable tweets from the selected OpenSources categories. Similarly, the tweets from the trusted MediaBias sources were assigned the label 1. Since we have the information for each account, we grouped the tweets by account id. Finally, a score can be assigned to each account based on the sum of the scores of all post from that account. This way, we not only have a data extraction workflow for distant annotation of reliable and unreliable social media posts but also for the accounts that disseminate them. In addition, several metrics can be developed and explored based on the extracted knowledge (Section 4.1).

3.1.2 Datasets

With the methodology presented in this chapter, two different datasets were created. The first dataset is tweet-related meaning that each entry corresponds to a tweet. In addition to the additional information we extracted from the Twitter API, we added two different labels. The first is the original label from OpenSources or MediaBias/FactCheck. The second is a binary label: "unreliable" if the queried link is from OpenSources and "reliable" if the queried link is from MediaBias/FactCheck. These data and labels are the basis for the work developed in Chapter 5. The original data can also be aggregated and transformed into an account dataset, by grouping reliable and unreliable posts by account. In addition, a numeric and binary label can be assigned to each account. The numeric label is simply the sum of the reliable (+1) and unreliable (-1) posts from each account while the binary label transforms that score in the "reliable" (account score > 0) or "unreliable" (account score <0). Since the work presented in this thesis was developed in a 4-year time span, the data mentioned in the subsequent chapters are subsets of the entirety of the extracted data. Not only did the subsets of data allow for continuous work over the years, but they are also an attempt to avoid some bias in the results presented. For example, the data used for the exploratory analysis described in the next section are independent of the data used for the content detection experiments conducted in Chapter 5. Similarly, the data used in Chapter 4 are a more complete version of the data presented in Section 3.2.2.The full datasets are available in this project Github repository [3].

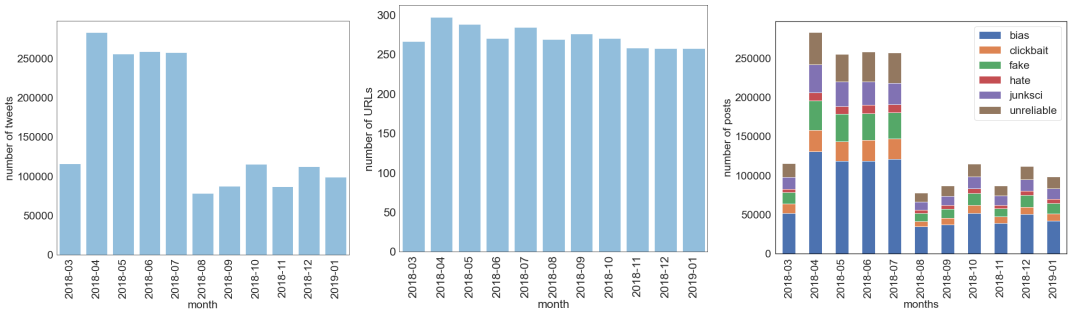
3.2 Exploratory Analysis

In this section, we present a preliminary exploratory analysis of the data. As mentioned earlier, for the sake of accurate representation and avoiding bias of the work conducted in the following chapters, this analysis uses the oldest data retrieved, corresponding to an interval of time of 10 months: from 20th March 2018 to 28th January 2019.

We divided our exploratory analysis into three main groups:

- a longitudinal analysis of unreliable content to understand the characteristics and patterns of this type of posting on Twitter using 10 months of data
- analysis of unreliable accounts to understand the dynamics of unreliable content disseminators
- a comparative analysis where we focus on the distinction between unreliable and reliable posts

[3] `https://github.com/nrguimaraes/PHDDatasets`



(a) Volume of Tweets *(b)* Unique Sources *(c)* Unique Classification

Figure 3.3: Volume of tweets, unique sources and unique classifications aggregated by month.

In the following subsection, we describe the different analyses conducted as well as the motivation behind each one.

3.2.1 Longitudinal Analysis

We begin our exploratory analysis by examining the characterization of unreliable tweets from a longitudinal perspective. Our goal with this analysis is twofold. First, to assess the robustness and quality of the retrieved data concerning diversification of labels and sources. In other words, if the tweets are only extracted from a very limited number of sources or labels, this may bias our analysis and limit the performance of our experiments to very specific scenarios. Second, to understand some common patterns in unreliable detection, and compare the current research with the literature.

The data analyzed in this subsection concerns unreliable posts retrieved between March 2018 and January 2019. We begin by presenting the volume of tweets extracted (Figure 3.3a), the number of unique sources (Figure 3.3b) and unique classifications by month (Figure 3.3c).

Regarding the volume of tweets, we can observe that April, May, and June have a larger number of unreliable posts captured than in the remaining months. However, the reasons for this increase are related to external constraints that occurred before and after this period of time. More specifically, since our retrieval process started on 20th of March, the number of retrieved tweets in this month is smaller. On the other hand, external factors related to the server where the crawler procedure was running affected the extraction process in the following months (e.g. lack of internet connectivity).

Regarding the different number of sources, we can see it varies between 250 and 300, from the total of 522 available sources. We hypothesize that the main reasons for this difference are related to the following factors:

- **Some sources presented in OpenSources do not post in a regular basis.** This

implies that their relevance is low, consequently affecting the spread of links from this source in Twitter.

- **The removal of the websites throughout the extraction process.** By manually visiting a sample of the websites listed in OpenSources, we found that some of them were unavailable which consequently affects the distribution of their content in Twitter.
- **The actions taken by Twitter to mitigate misinformation.** With Twitter focusing on fighting misinformation, we hypothesize that content from a large number of these websites is being continuously removed from the platform [64, 231].

Regarding the distribution of tweets by class, it is the bias label that achieves the largest share of tweets retrieved each month. This is partially due to the largest number of "bias" websites included in OpenSources. We also hypothesize that tweets containing links to extremely biased websites are less likely to be blocked or removed than tweets that spread false information. In fact, by examining the volume of tweets extracted each month, we can observe that the most frequent class is the "bias" class despite the fact that the "fake" class has approximately 100 more sources available. This reinforces the previously stated hypothesis on the internal mechanisms of Twitter to fight misinformation. Given the importance of removing "fake news" from the platform, it is likely that a greater effort was made to remove tweets containing links from this category. We can also observe that the distribution of classes is approximately the same over the entire period analyzed.

From these analyses, we can draw some conclusions regarding our extraction methodology. First, diversification of sources is ensured by the current method which means that when using this data to build unreliable detection models, it is less likely that these will be biased to tweets that spread content from a particular source. In other words, if the number of sources was low, the detection models could learn how to distinguish tweets that spread content from that source instead of tweets that spread unreliable content. Nonetheless, the volume of tweets for each source may vary which may affect the models. However, this imbalance is more difficult to handle, as balancing tweets by class and source would result in a smaller dataset. Regarding classification, it is clear that there is some imbalance in the volume of tweets extracted. However, the relationship between the number of sources and tweets is not linear and thus, one could argue that some of the classes are more targeted by Twitter's internal mechanisms than others. Therefore, although we acknowledge that the extraction process favors the bias class, we choose to maintain this imbalance because, similar to the previous example, undersampling to the smaller available class would result in a significantly lower number of tweets per month.

The second part of this section focuses on text analysis. More specifically, on the temporal dynamics of some of the key elements that have been studied in the literature on unreliable content, such as sentiment, emotion, entities, and hashtags. Our goal with this analysis is

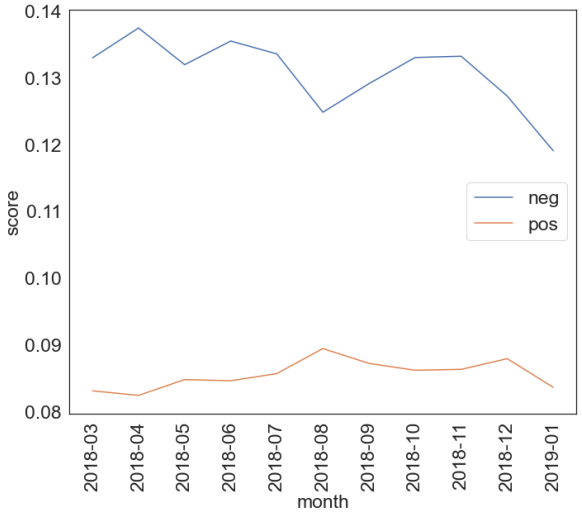


Figure 3.4: Negative and positive sentiment scores of unreliable tweet averaged by month.

to understand whether the introduction of new topics over time affects the overall emotions and the sentiment portrayed in unreliable tweets. Similarly, we would like to understand how these topics reflect on the entities detected and particularly in hashtags. Well-known movements such as #MeToo and #BLM (Black Lives Matter) have used hashtags to aggregate topics in the social network ecosystem. This research intends to examine whether the same is happening with unreliable content and its more specific causes/movements (e.g. the "lock her up" movement that sought to jail presidential candidate Hillary Clinton).

We used Vader [116] to examine the positive and negative sentiment of tweets and NRC Emotion Lexicon [154] for emotion detection. We assess these in each unreliable tweet extracted and average the results by month. Figures 3.4 and 3.5 show the sentiment and emotion scores (respectively) achieved in the time interval considered.

The average negative score is the highest in all the data which is consistent with the majority of the related literature that finds a correlation between negative sentiment scores and fake news or unreliable content [262]. Regarding emotions, they can be grouped into three different tiers. The first tier is formed only by the highest-scoring emotion trust. Its score remains stable, slightly fluctuating between 0.26 and 0.27. The second tier is composed of "surprise", "anticipation", and "fear" with these emotions' scores changing their position over time. Most noticeable, the sudden decay of the surprise score and the rise of anticipation score around August 2018. Several events at that time could lead to such a rise such as the removal of conspiracy theorist Alex Jones from all major streaming platforms [106], the assassination attempt on the president of Venezuela (Nicolas Maduro) [18], and the nomination of Brett Kavanaugh to the United States Supreme Court and the controversy that followed [109]. However, it is difficult to assess if these were the main triggers of such changes.

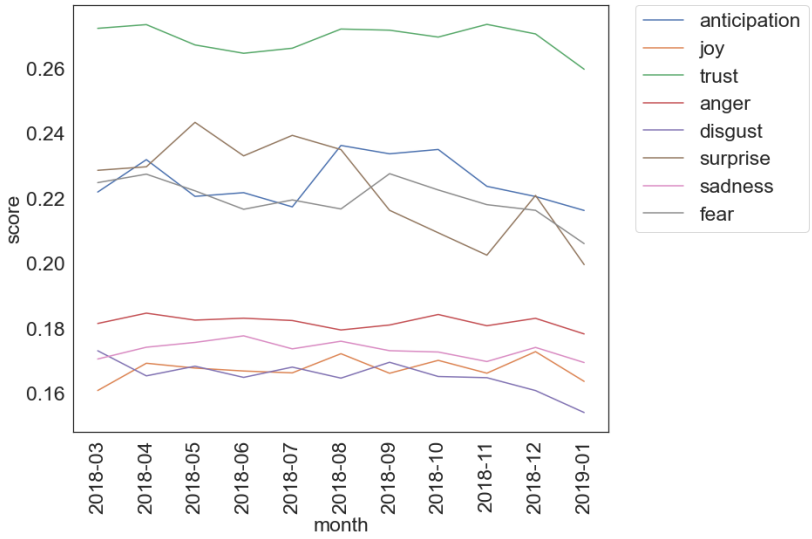


Figure 3.5: Emotions scores of unreliable tweet averaged by month

The third tier and lowest tier of emotions are anger, sadness, joy, and disgust. The majority of these emotions remain steady through time, with only joy and disgust briefly changing their position in the ranking.

The results of this emotion analysis differ from the results of some works. For example, in [247], responses to unreliable content achieve higher scores in surprise and disgust while in true news replies express trust, anticipation, joy, and sadness.

To better understand the dynamics of the topics discussed and disseminated over time, our next analysis focuses on the entities and hashtags disseminated in unreliable tweets.

We begin by examining the 20 most disseminated hashtags in each month, shown in Figure 3.6 using wordclouds, where the size of the word is proportional to its frequency in tweets.

Some interesting information is presented in the wordclouds. First, some hashtags remain consistently in the upper part of the top 20, with MAGA (an acronym for the slogan "Make America Great Again") being the most prevalent. Second, there is lean towards the political domain. Some examples include MAGA, Kavanaugh, Tea Party and Trump. Another interesting hashtag is PJNET which stands for Patriot Journalist Network, a Twitter group responsible for coordinating tweets and propagating "hyperbolic or false claims" [128]. There is also the mention of other groups such as TCOT and Red Nation Rising. In addition (and as expected) some hashtags refer to topics that were trending in the news at the time such as Trump, Syria, Israel and Russia. Therefore, we can conclude that there is a mix of hashtags related to relevant/newsworthy topics, with Twitter-specific groups and propaganda. In addition, we can notice that some hashtags are constant over time while others are derived from specific events from the time interval they were extracted (e.g. MarchForOurLives).

Figure 3.6: Wordclouds for the top 20 most frequently used hashtags in unreliable tweets captured by month.

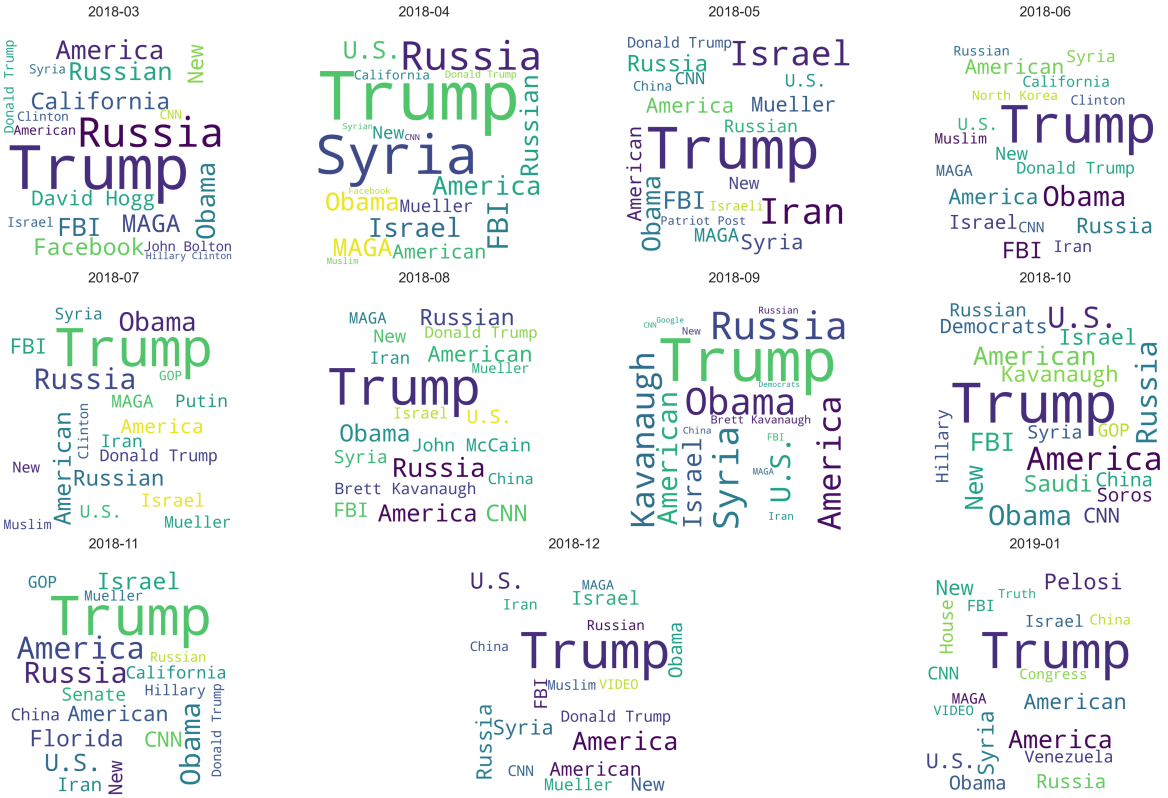


Figure 3.7: Wordclouds for the top 20 most frequently used entities in unreliable tweets captured by month

We can hypothesize that, on the one hand, these hashtags lead to user engagement by using interesting and relevant topics and on the other hand, they try to associate the publications with a specific group or page in an attempt to reach a larger audience for this specific type of content.

We conduct a similar analysis using the NLTK tool [25] to perform named-entity recognition on tweets. The 11 monthly-based entities wordclouds are presented in Figure 3.7. Similar to the previous analysis, some entities remain relevant throughout the time interval considered. Furthermore, the entities frequently detected can be strongly associated with the political domain. For example, Trump, Obama, Mueller, Kavanaugh, and Hillary are individuals who are associated with the United States Government and the political landscape. In addition, various locations and organizations are also frequently detected, such as Russia, U.S., Israel, and Syria. While some of these entities are constantly present in the months analyzed, others are associated with events that take place at specific times, triggering the spread of unreliable content (e.g., the controversial nomination of Brett Kavanaugh to the United States Supreme Court [32]). This analysis shows evidence that unreliable content on Twitter discusses topics that are constant over time such as Trump, Obama, and Russia but also includes topics that are trends or associated with events that occur at specific

intervals, such as the previously mentioned nomination of Brett Kavanaugh and the death of John McCain [66]. Consequently, when developing unreliable content detection models, these emerging topics could lead to a drop in performance. Thus, these results reinforce the importance of evaluating these models in a long-term scenario where the train and test data are chronologically ordered.

We proceed to analyze the accounts associated with the extracted unreliable tweets.

3.2.2 Account Analysis

In this section, we focus on analyzing some characteristics of accounts that disseminate unreliable content and the types of unreliable content that are disseminated. We formulate the following 4 research questions (RQ).

RQA1: Do the most active spreaders propagate content from a single or multiple unreliable classes? Given the different types of unreliable content on social networks, our first research question seeks to investigate whether the most frequent accounts focus on disseminating single (i.e. one type of unreliable content per account) or multi-class content.

RQA2: What types of unreliable content are most frequently disseminated simultaneously through a single account? This research question is a follow up to RQ1 and investigates which types of content are often diffused together (i.e by the same online social network account)

RQA3: Do verified accounts spread unreliable content? A verified account is one that Twitter confirms belongs to a specific relevant entity, such as a company, celebrity, or politician [234]. These accounts tend to be trusted due to the authority associated with them and are followed (in some cases) by millions of people. Thus, an important question is if these verified accounts are spreading unreliable content.

RQA4: What is the current status of the unreliable accounts retrieved? Are accounts spreading unreliable content being removed/suspended? Twitter has been actively fighting the spread of unreliable content on its platform. However, this task is not trivial and unreliable accounts are still present. With this RQ we intend to investigate the percentage of accounts that have been captured by our method and are still active, as well as the accounts that have been removed or suspended from the platform.

With respect to RQA1, we listed the 30 most frequent spreaders identified (i.e. the accounts with more posts in the data considered). Figure 3.8 shows the number of posts of each type of unreliable content for each account. In the top 30, there is a combination of accounts that focus on one type of content with accounts that diversify in the type of unreliable posts diffuse. The account with the most posts not only has a large volume of tweets captured, but

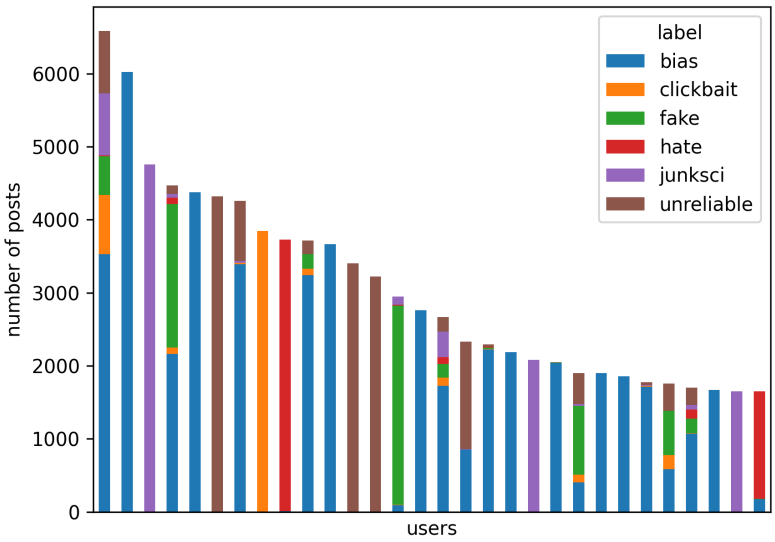


Figure 3.8: The number of posts discriminated by type of unreliable content for the top 30 accounts.

also a large diversity in the types of unreliable content it posts (5 different classes) compared to the remaining accounts in the top 30. There is also a large proportion of accounts (more than half) whose focus is on a single class, with "bias" content being the most prevalent. This factor, as well as the high presence of extremely biased content in multi-class accounts can be explained by the imbalanced number of sources and consequently, a larger number of tweets captured. Thus, to answer our first research question RQA1, we can conclude that among the top 30 spreaders there is a balance between accounts that disseminate content from a single class with accounts that disseminate tweets associated with multiple classes.

In RQA2, we extend the previous research question to include every account captured and analyze the classes that tend to co-occur more frequently in unreliable accounts. For this purpose, we count the number of accounts that disseminate posts from each pair of classes. Therefore, if an account disseminates content from 3 classes it will appear at least in three pairs, so accounts can overlap in classification. The co-occurrence of classes is presented in Figure 3.9.

The pair (bias, unreliable) is the one that reaches the highest number of accounts. In other words, 41741 accounts contain at least one unreliable and one biased post. Similarly (bias, fake) and (bias, clickbait) are also posts that tend to occur frequently in the same account. It is important to highlight that the high volume of bias sources and tweets are likely to influence the presence of this class in the top pairs and thus the results should be interpreted accordingly. Other classes that frequently occur together are unreliable and clickbait (22030),

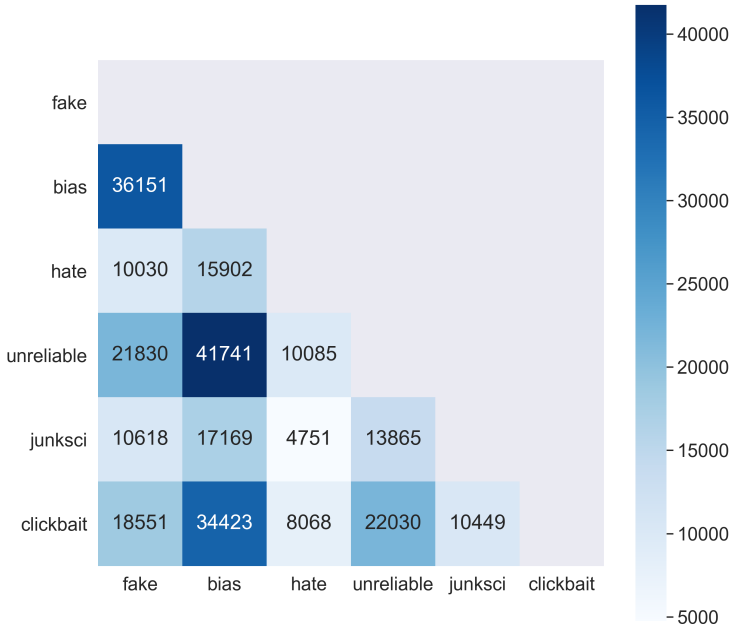


Figure 3.9: Number of accounts that diffuse content in each pair of classes.

and unreliable and fake (21830).

The previous two analyses show evidence that accounts often disseminate content from different classes of unreliable content and thus limiting account detection to a specific type of content can often exclude or limit the extent of the problem. Based on these results, detecting unreliable accounts as a binary classification problem provides a way to capture not only accounts that disseminate a single type of unreliable content but also accounts that disseminate multiple types of unreliable content.

Finally, to answer RQA3 and RQA4, the verification status of all accounts was extracted using the Twitter API. Due to changes in the crawlers and APIs, some of the verification statuses were not extracted during the time period analyzed. Thus, to standardize the data, for this analysis we extract the information at the time of writing this chapter (November 2021). Consequently, some accounts may have gained or lost their verification status after the extraction process. Figure 3.10 shows the percentage of verified and not verified accounts, as well as the status of accounts whose verification status was unknown due to their suspension or removal.

Regarding RQA3, we can observe that approximately 1% of the unreliable accounts extracted

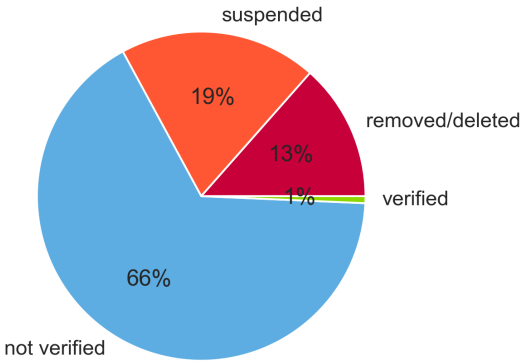


Figure 3.10: Unreliable accounts' status at the time of verification (November 2021).

during the considered time period are verified. This result is similar to the analysis previously performed and published in samples of the current dataset wherein 72000 unreliable accounts, 407 (0.6%) were verified [94]. This analysis provides the important insight that verified accounts are much less likely to disseminate unreliable information. Nevertheless, it is important to highlight that verified accounts also represent only a small percentage of all Twitter users, as only 360K [84] of the 211 million active [239] accounts are verified.

Regarding RQA4, the high percentage of accounts that have been removed or suspended (32%) provides some evidence on the efforts of Twitter to tackle unreliable content. While we are aware that the removal/deletion of some accounts may have been initiated by the account owner, due to the continuous Twitter press releases regarding the removal of the accounts and the nature of these accounts [194, 3], it is reasonable to assume that the majority of this percentage was removed due to their malicious activity in the social network ecosystem. On the other hand, the suspension of Twitter accounts are frequently associated with non-compliance with Twitter rules and policies. Consequently, although these can be temporary, the larger percentage of suspended accounts shows once again the efforts of Twitter to crack down on accounts that disseminate unreliable content. Nonetheless, considering that posts from these accounts were retrieved in 2018 and 67% of these accounts are still active in 2020, it is clear how difficult this task is.

3.2.3 Comparative Analysis

To better understand the differences between reliable and unreliable tweets, we focus on analyzing a sample of both types over a smaller time interval. Hence, in this section,

we examine reliable and unreliable tweets from April 2018. In addition, we also use the binary labels associated with each tweet, meaning that we focus on distinguishing between reliable and unreliable content. Similar to the current literature, we explore some of the characteristics that might help to distinguish reliable and unreliable tweets such as sentiment, emotion, entities, and hashtags.

Based on the previous literature, we formulate three research questions to assist us with the analysis of useful indicators of reliable and unreliable tweets.

- **RQ C1: Does sentiment differ between unreliable and reliable tweets and tweets?** An important factor suggested by the authors in [62] is that news headlines are more attractive when their sentiment is extreme. The same analysis was later conducted with "fake news" [217] with the conclusion that this type of content presents more negative sentiment than mainstream news outlets. Therefore, it is important to analyze whether such a pattern can also be found in reliable and unreliable tweets.

- **RQ C2: Does emotional tone differ between unreliable and reliable tweets?** Vosoughi et al. [247] suggested that false rumors elicit more surprise and disgust, while true news elicit replies with sadness, anticipation, joy, and trust. Given that this analysis was conducted on responses to true news stories and false rumors, we intend to examine whether the same emotions are present in reliable and unreliable tweets. Although a similar breakdown has already been done in Section 3.2.1, in this analysis we focus on the difference between the content of both types.

- **RQ C3: Do unreliable and reliable tweets discuss the same topics?** The authors in [242] have concluded that traditional online media outlets appear to be responsible for the fake news agenda since the content of traditional media makes users more attentive to all online content regarding the topics discussed in this. With this RQ we want to examine whether this is also the case for tweets. In other words, we intend to investigate if entities and hashtags in reliable and unreliable posts overlap, thus providing clues to the similarity/divergence of the topics discussed.

- **RQ C4: Do unreliable tweets contain more hashtags, mentions, and entities?** The majority of accounts posting unreliable tweets have the goal of spreading them across the network. Thus, we hypothesize that unreliable tweets will contain more social engagement features such as mentions and hashtags. In addition, the frequency of these features as well as the different types of entities could provide important evidence on the distinction between reliable and unreliable tweets.

We start by evaluating the sentiment of reliable and unreliable tweets in the time period considered. Similar to what was done previously, we evaluate the sentiment scores using

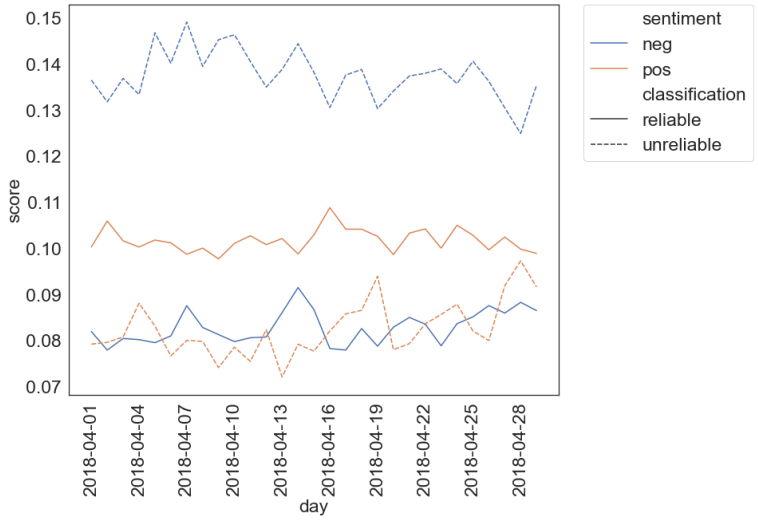


Figure 3.11: Sentiment score for reliable and unreliable tweets aggregated by day

Vader and aggregate them by day for each of the classes. The results are showed in Figure 3.11.

A clear predominance of negative sentiment is visible in unreliable tweets with the score associated maintaining the highest value throughout the period examined. The highest value for positive sentiment score is obtained from reliable tweets, preserving a constant value between 0.10 and 0.11. Finally, the lowest sentiment scores are obtained in the negative class in reliable tweets and in the positive class in unreliable tweets. These scores vary between 0.08 and 0.09 with the highest-scoring class changing in some cases. Hence, to answer our RQ C1 and similar to the longitudinal analysis conducted previously, these findings support previous literature that correlates negative sentiment with unreliable or fake news content [262].

With respect to the emotions presented on reliable and unreliable tweets and to answer RQC2, we use the NRC Emotion Lexicon [154] to assign emotion scores to the reliable and unreliable tweets, averaging the results by day. Figure 3.12 compares the scores for each emotion in each class.

The current data presents a very similar pattern in emotions in reliable and unreliable tweets. The emotion "trust" is the one that achieves the highest score in both classes. We assume that both reliable and unreliable tweets are perceived as trustworthy. Reliable tweets because they come from reputable sources and unreliable tweets due to their intention to be perceived as trustworthy so that the information they spread is believed by Twitter users. Subsequent emotions scores are the ones where reliable and unreliable tweets begin to contrast. While reliable tweets show a clear difference between the following two top emotions (anticipation and fear) over time, unreliable tweets present a blend of these two as well as the emotion

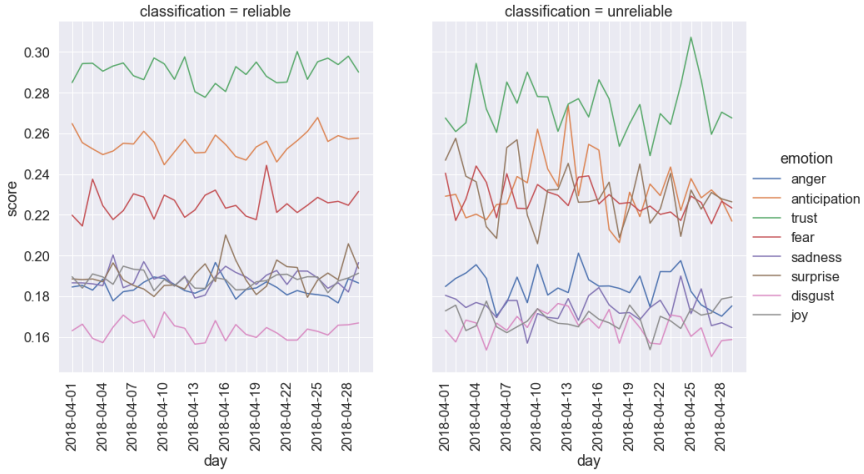


Figure 3.12: Emotions scores for reliable and unreliable tweets aggregated by day

surprise, with a more chaotic distribution over time where the highest scoring emotion varies over the defined data interval. The lowest scoring emotions in the unreliable data are anger, joy, sadness, and disgust with a similar pattern to the previous ones (i.e. the variation of the highest scoring emotion over time). Similarly, reliable tweets present the same pattern for surprise, anger, sadness, and joy but present a clear lowest scoring emotion (disgust) whose values are always lower than those of the remaining emotions. These results go against the results in [246] which suggest that false rumors elicit replies with greater surprise and greater disgust while in real news replies tend to be associated with sadness, anticipation, joy, and trust. Therefore, to answer RQ C2, there are some differences in the emotion scores of reliable and unreliable tweets, with the top four emotions in unreliable tweets showing a more chaotic pattern through time while in reliable tweets, each top emotion score is constant throughout the time period analyzed.

We proceed to investigate if unreliable and reliable tweets cover the same topics by analyzing the top hashtags and entities mentioned. Figure 3.13 presents the wordcloud for the top 50 entities and hashtags in reliable and unreliable tweets.

In the entities wordclouds, terms such as Trump, Syria, US/U.S., Russia are frequent in both classes with a high prevalence and similar proportions, suggesting that some relevant topics are frequently discussed in both reliable and unreliable tweets. There are also entities that are more likely to appear in one of the categories. As an example, "Obama" occurs more in unreliable tweets than in reliable ones, while it is the other way around for "Facebook". We hypothesize that this type of contrast emerges due to the interest of the topics being discussed to the readers/consumers of each type of content. For example, due to the Cambridge Analytica Scandal on Facebook, it makes more sense that this topic is discussed in reliable content and ignored by unreliable content due to the implications it had on the propagation of fake news [134, 17].

(a) Entities from reliable tweets

(b) Entities from unreliable tweets

(c) Hashtags from reliable tweets

(d) Hashtags from unreliable tweets

Figure 3.13: Wordclouds of entities and hashtags regarding reliable and unreliable tweets.

The overlap between reliable and unreliable tweets hashtags is less frequent and is more specific to each class. Unreliable tweets hashtags are very similar to the analysis conducted previously in Section 3.2.1 with a combination of news topics (e.g. Syria, Israel and Russia) with hashtags associated with the spread of propaganda and unreliable content (e.g. pjnet, qanon, rednationrising, maga). For reliable tweets, the top hashtags are more related to the topics being discussed. In addition, these topics are broader than the ones presented in unreliable tweets, ranging from politics (#Trump,#US) to music and TV shows (#BTS, #GOT7,#KPOP). It is also worth mentioning that in some cases, media-specific hashtags may also appear (e.g. #SkimmLife which is associated with the news medium TheSkimm).

In summary, to answer RQC3, we find that the most frequent entities mentioned in reliable

and unreliable tweets overlap. Nevertheless, some additional entities also appear on both sides of the spectrum but in different proportions. Consequently, some of the topics discussed are very similar in both classes. However, analogous to the previous analysis (Section 3.2.1) the focus on unreliable tweets seems to be on account engagement, which is achieved through the use of hashtags associated with groups whose interests may be suited to the users of that particular content. However, in reliable tweets, the top hashtags are derived from the topics being discussed and less on the self-promotion of groups or news mediums. Thus, reliable tweets use hashtags to point towards specific topics while unreliable tweets focus more on using hashtags for propaganda of specific groups.

Our final research question concerns the frequency of entities, mentions, and hashtags and how reliable and unreliable tweets differ in these indicators. To this end, in Table 3.3 we summarize the data regarding the different entities - Persons, Organizations, Locations, Geopolitical Entities (GPE), Geographical Social-Political (GSP), and Facility (e.g. Washington Monument and Stonehenge) - as well as hashtags and mentions. It is important to highlight that the mean and standard deviation presented is referring only to tweets that contain at least one of the indicators analyzed. In other words, before the average and standard deviation is computed to a certain indicator, all tweets without that indicator are removed from the analysis. This process helps to highlight the difference in the number of entities/hashtags/mentions per tweet containing that feature, avoiding that the tweets without it affect the metrics. To complement the analysis, we also reported the percentage of tweets that contain at least one of the analyzed indicators.

Table 3.3: Summarization of the frequency of different entities in reliable and unreliable tweets and the percentage of tweets that contain at least one of the indicators analyzed.

	Hashtags	Mentions	GPE	Organization	Person	Facility	GSP	Location
	Unreliable							
mean	**0.46**	0.92	0.45	0.98	1.07	0.02	0.06	0.01
std	**1.25**	**1.52**	**0.71**	**1.15**	**1.08**	**0.15**	**0.27**	0.10
% of tweets	0.20	0.69	**0.35**	**0.59**	**0.63**	**0.02**	**0.06**	**0.01**
	Reliable							
mean	0.40	**1.11**	**1.24**	**1.41**	**1.42**	**1.01**	**1.06**	**1.01**
std	0.99	1.12	0.55	0.76	0.72	0.13	0.25	**0.12**
% of tweets	**0.23**	**0.81**	0.29	0.41	0.43	0.01	0.02	**0.01**

The summarized results show a higher average for entities and mentions in reliable tweets. In addition, the standard deviation is lower for reliable tweets (except for locations) indicating a possible more distributed number of entities in this class. Some factors that likely contribute to these results are the fact that some of these tweets are more informative and factual and less opinion-based (since they are extracted based on news sources).

The percentage of tweets containing at least one occurrence is higher in unreliable tweets for all entities analyzed. This is reversed for hashtags and mentions, as reliable tweets have a higher percentage for these indicators.

Thus, the evidence provided in Table 3.3 does not support what we hypothesize in RQ C4. The number of hashtags is higher on average in unreliable tweets. However, the number of mentions is higher in posts of the reliable class. In addition, the presence of hashtags and mentions is more likely in reliable tweets than in unreliable tweets. The average number of entities in a reliable tweet for each category is also higher than the average number in unreliable tweets, with a lower standard deviation.

3.3 Conclusions

In this chapter, we have presented the extraction methodology adopted and modified from the literature, to retrieve and annotate tweets in terms of the type of content they disseminate. We also described how this methodology can be used to develop tweet and account-based datasets. An exploratory analysis is then conducted to compare the extracted data with the current literature in terms of some key characteristics of reliable and unreliable content. The results are inconsistent, as some of the research supports previous literature while other does not. We hypothesize that some of these inconsistencies may result from the different data where the analyses were conducted. For example, although the emotion classification tool used in this chapter was the same as the one used in [247], the analysis in the latter was conducted only in replies to false news.

In the next chapter, we address on the problem of detecting reliable and unreliable accounts in social networks, approaching the problem in a more realistic scenario, where knowledge-based and prediction-based strategies can be used.

Chapter 4

Towards a Pragmatic Detection of Unreliable Accounts

Concerning the task of detecting unreliable accounts on social networks, the current state of the art has mainly focused on the bot detection task. However, as we discussed earlier, there is evidence that human-operated accounts are also responsible for the spreading of unreliable content. Moreover, an approach considering a more pragmatic scenario is missing, especially in the context of the information available in individual accounts. Hence, in this chapter, we address the problem from an unreliable vs reliable accounts perspective.

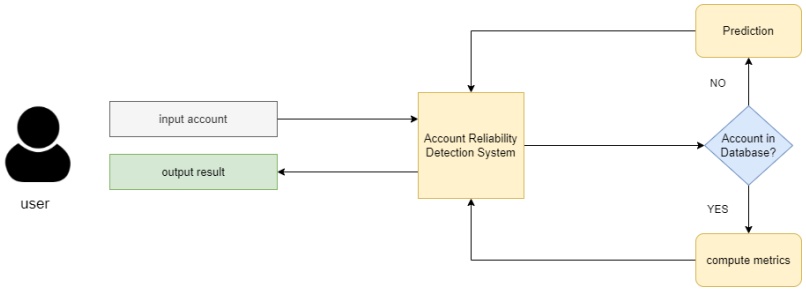


Figure 4.1: A experimental account reliability classification system based on the work conducted in this chapter.

Furthermore, taking a more pragmatic approach, we focus on two different scenarios that can be used as a basis for the development of unreliable detection systems. First, the development of metrics based on previously extracted knowledge. More specifically, using the methodology detailed in Chapter 3, a database can be built with unreliable and reliable accounts. Consequently, if an unreliable account detection system already has information stored about the account, this knowledge can be used to evaluate the reliability and influence of that account in the social network. The development of these metrics is presented in

Section 4.1. When such knowledge is not available, then a prediction about the reliability of that account must be made based on the information available online. In Section 4.2 a classification-based approach is presented based on the constraints of the information available for each account. A possible solution for an unreliable account detection system that incorporates the work presented in this chapter is illustrated in Figure 4.1.

4.1 Knowledge-Based Approach

Based on the data extraction methodology and exploratory analysis conducted in the previous chapter, we develop metrics to analyze Twitter accounts based on the number of unreliable and reliable posts captured as well as some additional features such as the social feedback of the posts and account information. Specifically, the following variables will be used to compute the metrics presented in this section.

- account creation date
- account number of followers
- account verification status
- post publication date
- post retweet count
- post favorite count

In addition, these metrics can also be used to classify accounts based on binary classes (reliable/unreliable) or in multiple classes (e.g. "hate", "false", "least bias", "pro-science"). We will discuss some case scenarios of using these metrics with the binary and later on with the multiclass variant.

We begin by introducing the first metric which serves as the basis for the remaining ones. This metric is based on the assumption that an account's behavior with respect to the content it publishes is time-independent and therefore an account's impact on the network is based solely on the number of tweets it publishes/shares. Thus for an account a and a class c, $pcount_{a,c}$ can be defined as the total number of posts from account a in the class c.

This first metric can be summarized into a post count and has already been discussed in Chapter 3. Nevertheless, this metric provides the basis for the classification of accounts. It is important to highlight that the main difference between this work and similar social network metrics in the literature is the knowledge-based component that was extracted and stored using the methodology described previously. This allows for a quick categorization

of accounts and stores the information needed for the computation of these metrics for later use in an unreliable account detection system.

The second metric is time-dependent (i.e. the older the post, the less impact it has on the reliability of the account) and looks at the behavior of the account, ignoring any output variables such as its connections or the social feedback obtained in its posts. We use the following notation to simplify the formulation. Each post i published by account a and annotated with a classification c, can be defined by a tuple $(t_{i,a,c}, f_{i,a,c}, r_{i,a,c})$ where $i \in [0, n_{a,c}]$, $t_{i,a,c}$ is the age of the post (in months- derived from the post's publication date), $f_{i,a,c}$ the number of accounts that "favorited" this post and $r_{i,a,c}$ is the number of retweets. In addition, we also used the age (in months) of the account (AGE_a).

The equation to compute the behavior (BEH) of an account for each class c is:

$$\text{BEH}_{a,c} = \frac{\sum_{i=1}^{n_{a,c}} \frac{1}{t_{i,a,c}}}{\text{AGE}_a} \tag{4.1}$$

Let us elaborate on the intuition behind this equation. Using the multiplicative inverse for the post's age allows for a linear decrease in the influence of that post in the assessment of the account behavior metric. Let us consider two different examples. The first is a bot account that was actively disseminating information prior to the 2016 U.S. presidential election ($account_1$). The second is a human-operated account that shares extremely biased content during the Covid-19 pandemic ($account_2$). It is reasonable to assume that the latter account should achieve a higher score because the post is more recent. In fact, in a simulated environment where $account_1$ propagates 50 posts from September to November of 2016 and an account that posts 10 posts in the last 2 months (May and June 2020), the results for the multiplicative inverse are 1.16 and 4.83 respectively. The sum of the multiplicative inverse of the age of each post allows us to quantify each post individually taking into account the respective time differences.

The next step in the metric is to divide the sum of all posts' inverse age from account a by the age of the account (AGE_a). There are two main reasons to reduce the effect of older accounts. First, the registration date plays an important role in most of the works concerning unreliable information [29, 254] with some works highlighting the importance of this feature [256, 38]. Secondly, due to the ongoing efforts of Twitter to remove bot accounts and accounts that disseminate misinformation [49, 145], it is plausible that accounts with a long history and constant propagation would be captured by the social network's internal algorithms. Once again, let us consider two example accounts: $account_3$ was created in 2016 while $account_4$ was registered in June 2020. Account 3 is a human-operated account that has published unreliable information 10 to 20 times in its lifetime, but has recently moved away from such content. On the other hand, $account_4$ has published 5 tweets containing

unreliable content. Due to the recent creation date of $account_4$, the penalty from the behavior metric would be higher than for $account_3$. It is reasonable to assume that a recently created account that propagates unreliable content in the first few months may be a bot or even an unreliable account that will continue to engage in the same behavior and therefore the unreliable metrics proposed should assign a higher value to it.

The third and final metric is the impact (IMP) metric which combines behavior and influence of an account on Twitter. Influence metrics on Twitter have been thoroughly proposed in social network studies [191] . For example, popularity metrics weigh the reach of account popularity based on its close connections. Examples of these metrics can be the in-degree measure (using followers and followees count) [100] and the Twitter Followers-Followees ratio [24]. Nevertheless, it is the metrics that allow quantifying the influence of a user/account that are largely studied. Adaptations of the Closeness (size of the shortest path from a node to every other) and Betweenness (number of shortest paths that pass through the node) were proposed in a social network scenario. Furthermore, several authors have proposed variations of the Page Rank algorithm applied to social networks [191, 258].

However, when incorporating these metrics in a real-world system, they must provide fast output and avoid heavy computation. Therefore, influence metrics that require knowledge of the close network (such as the closeness, betweenness, h-index [188] or PageRank [249]) are not suitable for this purpose. Therefore, our influence metric relies solely on data that can be derived from the account information provided by the API. A metric that is more suitable for our goals is the information diffusion metric [165]. This metric is used to estimate the influence of an account in a topic by measuring the difference between the number of friends of an account that tweets on a topic before and the number of followers of the account that tweets on the same topic after. The metric also uses a logarithmic scale due to the possible differences between the number of followers (NFOLLOWS) and friends (NFRIENDS). However, this metric still relies on information from the connections to be computed and thus is subject to the limitations of the Twitter API. Metrics based only on the number of followers and friends (and therefore can be computed almost instantly) were also proposed. For example the Followers Rank [158] (presented in Equation 4.2) and Twitter Follower-Followee Ratio [24]. The first measure is the adaptation of the in-degree metric (where NFOLLOW and NFRIENDS represent the number of followers and number of friends of an account, respectively) and the second is self-explanatory.

$$\text{FollowersRank} = \frac{\text{NFOLLOW}}{\text{NFOLLOW} + \text{NFRIENDS}} \tag{4.2}$$

Nevertheless, these metrics have their limitations. The first is the disproportion that can exist between NFOLLOW and NFRIENDS. For example, at the time of writing [1], the official

[1] June 2020

Twitter account of the (now former) president of the United States (*@realDonaldTrump*) had 82.9 million followers but only 42 friends which can greatly affect the value of this metric compared to more balanced accounts.

Furthermore, we argue that in the specific domain of unreliable content, the verification status of an account can play an important role in the influence of that account, and thus it should be considered. Twitter assigns a verification status to accounts that are of public interest and authentic [237]. Therefore, if a verified account publishes unreliable posts, users might more easily believe that the content is reliable due to the account's authenticity and authority.

Due to the aforementioned reasons, we present the (INFLUENCE_a) metric, which can be measured using the following equation,

$$\text{INFLUENCE}_a = \log(\text{NFOLLOW}_a + 1) \times \alpha^{VER_a} \tag{4.3}$$

where VER_a refers to the verified status of the account (0 if it is not verified and 1 otherwise) and α user-defined variable with a value greater than one ($\alpha > 1$) that assigns the weight applied to the influence of an account when it is verified. We also restrain $NFOLLOW$ on a logarithmic scale because of the large discrepancy in the number of followers of Twitter accounts.

Another clear difference between our influence metric and other metrics presented in the literature is the exclusion of the number of friends (or followees). The reason behind this decision is the specific domain in which we wish to apply this metric. When measuring the influence in terms of the overall popularity of an account, a ratio between the number of followers and followees is essential since high influence accounts (belonging for example to musicians and actors) have a large discrepancy on these numbers due to the fans that follow these particular accounts. However, when assessing the impact of an account in spreading unreliable content, it is reasonable to assume that the influence it has on its followers should not be reduced by a high number of friends/followees ratio. For instance, if we consider two accounts with the same behavior and the same number of friends, but the first account follows many more accounts than the second, it is plausible that both accounts have the same reach in their close network. In other words, an unreliable account with a high number of followers should not have the same impact as an account with a high number of followees because the publications of the second one have a higher risk of being scattered. In addition, since that high disparity of followers and followees is usually associated with accounts of public interest, the verification status component in our metric can account for these particular cases.

Finally, combining influence and behavior, we characterize the impact of an account using the following equation:

$$\text{IMP}_{a,c} = \text{BEH}_{a,c} \times \text{INFLUENCE}_a \tag{4.4}$$

We can further complement the metrics based on information from social feedback such as retweets and favorites. Thus, we propose a variation of the behavior (Equation 4.5) and impact (4.6) metrics that takes into account the social feedback of tweets at the time of the retrieval.

$$\text{BEH\_}SF_{a,c} = \frac{\sum_{i=1}^{n_{a,c}} \frac{1+\log(1+f_{i,a,c})+\log(1+r_{i,a,c})}{t_{i,a,c}}}{\text{AGE}_a} \tag{4.5}$$

$$\text{IMP\_}SF_{a,c} = \text{BEH\_}SF_{a,c} \times \text{INFLUENCE}_a \tag{4.6}$$

By adding the number of retweets and favorites in each post to the behavior function we aim at a better characterization of the impact on the network. We use the number of retweets and favorites in individual logarithmic functions to highlight the difference between posts with a large number of only one type of social feedback from the ones that have a large number of both favorites and retweets. The value is smoothed by the age of the post and the age of the account for the reasons already mentioned.

However, there is also a limitation on the application of social feedback. Since we are trying to create a knowledge-based metric that relies on information that already exists in our database (thus avoiding calls to APIs or computationally expensive processes) we rely solely on the social feedback at the time of extraction. In other words: Over time, the number of retweets and favorites of each tweet may increase. Some measures have been taken to minimize the dynamic effect of some of these variables. As previously described, we update the values each time a duplicate tweet is found. Since the Twitter Search API is used, we can update a tweet within 7 days of its publication in the best case scenario. Second, when searching for tweets, we specify that the search should consist of a mix of popular and recent tweets by using the "mixed" parameter from the Twitter API [235]. This compromise allows us to capture both recent tweets that may not be engaging (and thus have no impact on the network, but are still relevant for broad characterization of Twitter accounts) and tweets that have already gained some traction on Twitter (and are therefore considered popular according to the API standards).

In the next section, we apply these metrics to the collected accounts. In addition, we compare the results obtained with a state-of-the-art bot detection system and provide reasoning on the usefulness of these metrics.

4.1.1 Case Studies

We present three case studies to evaluate the usefulness of our metrics. First, we analyze the most reliable and unreliable accounts captured (subsection 4.1.1.1). Second, we examine

how the metrics perform compared to a state-of-the-art bot detection method (subsection 4.1.1.2). Third, we investigate the current state of Twitter concerning unreliable information in trending and relevant topics by extracting a sample of publications from the APIs, and comparing their authors to the account knowledge acquired from the previous extraction process.

The data used in these case studies were extracted between July 2019 and July 2020. It includes over 4M tweets and more than 750k distinct accounts.

In Chapter 3, when analyzing a sample of unreliable accounts, we found that only a small percentage ($\approx 1\%$) were verified. In addition, verified accounts often have a large number of followers. Therefore, in these case studies, the value of alpha was set to 2 to avoid treating verified accounts as outliers in our IMP_{SF} metric but, on the other hand, to ensure that they could be distinguished from non-verified accounts. Defining $\alpha = 2$ means that if an account is verified, its influence value doubles.

4.1.1.1 Top Unreliable and Reliable Accounts

In this case study, we select the 5000 accounts with the highest $PCOUNT$ value for each class (reliable and unreliable). We chose this metric because 1) it is the baseline metric and 2) it allows us to examine how this baseline metric differs from the more complete metrics proposed (BEH and IMP). If one of the other metrics was chosen, other characteristics (such as the number of followers or the age of the account) might be less propitious to change and thus making the analyzed accounts more similar.

Unreliable Accounts Regarding unreliable accounts, 44 accounts were verified in our sample. The majority of these were Twitter accounts of the websites annotated in Open-Sources or entities associated with those websites (such as reporters or commentators). The non-verified accounts that have the highest IMP_{SF} and BEH_{SF} are bots that publish and disseminate extremely biased content. We select the top 5 for a more in-depth analysis. However, one of these accounts had been already suspended at the time of this analysis. Screenshots from the remaining 4 are presented in Figure 4.2.

The similarity of the accounts presented in Figure 4.2 as well as the absence of a personal profile picture/banner and lack of original publications, are clear signs that these are bot accounts. Three of the four accounts have a high number of followers (between 13000 and 25.8K). In addition, they have a recent registration date with the oldest account being registered less than two years earlier to the report of this analysis, and the other three being registered approximately a year ago. These factors combined with the number of posts retrieved justify the score they achieved. Furthermore, we argue that the score assigned

Figure 4.2: Characteristics of the accounts with the highest IMP_{SF}. The similarities of the accounts are visible and present all the indicators of bot accounts.

to these accounts is fair in the sense that these accounts exhibit bot behavior and pose a threat to the upcoming [2] 2020 United States elections due to the influence of their content on the social network. The last account shown in Figure 4.2 has a lower number of followers in comparison. However, its IMP_{SF} score is severely affected by the high number of posts captured and it is the account with the highest BEH_{SF} value in this sample. Hence, although its number of connections is low, it was recently created and systematically propagates unreliable content, making it also a potential problem in the network. When compared with the other analyzed accounts, this account has 2076 posts captured in our database while for the remaining the number of posts ranges from 407 to 737.

We proceed to manually analyze some random examples of accounts that have a IMP_{SF} value between 2000 and 4000 and a BEH_{SF} value between 250 and 500. Among the 20 accounts that were analyzed manually, there was a mix of bot and human-operated accounts. Although a high presence of bots can be seen, it is clear that some accounts exhibit human behavior. In some cases, these accounts feeds have a high proportion of retweets and a low

[2] regarding the timing of the analysis

proportion of original posts. In other cases, they have a large volume of original posts and a small proportion of posts linking to unreliable websites.

Finally, we manually analyzed some of the accounts with low unreliable scores in our metrics. We selected accounts with a BEH_{SF} score between 10 and 150 and IMP_{SF} between 40 and 100. Once again there is a combination of human-operated and bot accounts but with a lower influence score and an older registration date than the previous tier. A curious example is an account that at the time of extraction had a BEH_{SF} value of 121.93 and an IMP_{SF} of 0 (since it had no followers). This case perfectly illustrates that even if some accounts have a high propagation of unreliable content, their impact on the network can be limited.

In summary, by manually sampling some of the accounts captured we can provide some confidence in the effectiveness of the metrics in annotating high-impact accounts that frequently disseminate unreliable content (such as bots), as well as human-operated accounts (whose propagation frequency is lower). In addition, this analysis also provides evidence that the metrics distinguish accounts that have high impact on the social network from those that do not.

Reliable Accounts We shift our analysis to reliable accounts and proceed to inspect a sample of cases where a high IMP_{SF} value is obtained. First, the number of verified accounts is higher than in the unreliable domain, with 221 of the accounts achieving this status. Again, a large number of these verified accounts are the official Twitter accounts of the websites presented in MBFC. Second, the values of the metrics IMP_{SF} and BEH_{SF} are lower than the ones obtained for the unreliable accounts. We hypothesize that due to the highest number of human-operated accounts, the feed of these accounts is more diversified with conversational threads and fewer publications with links to reliable websites.

We proceed to manually analyze some of the accounts. One of the first observations that illustrates the need for these metrics to complement current bot detection systems is that the account with the highest behavior score is a bot that retweets news from multiple reliable sources (@world_news_eng). This account was created recently [3] (January 2020). Hence, it does not have a high number of followers and consequently does not have a high IMP_{SF} score. It is also important to mention that similarly to the unreliable accounts, there are some accounts that were removed at the time of this analysis. Furthermore, some manually analyzed accounts have moderated biased opinions. However, their information and opinions are based on information from reliable websites/sources.

Nevertheless, comparing the metrics in both classes, it is clear that the impact of accounts that disseminate unreliable information on Twitter outweighs the impact of accounts that disseminate reliable content.

[3] regarding the timing of the analysis

4.1.1.2 Botometer vs Reputation Metrics

To understand how the developed metrics compare to the traditional state-of-the-art bot detection systems, we analyzed and evaluated a sample of the accounts extracted using Botometer. This tool, formerly known as BotOrNot, is one of the state of the art systems for bot detection [243]. Although the metrics presented in this paper aim to detect unreliable accounts (bots and human-operated), we compare the scores assigned by Botometer with those computed by our metrics to 1) understand the similarities/differences between the two scores 2) reinforce the necessity of the metric proposed in this chapter.

For this experiment, we used the "universal" Botometer score to determine if an account is a bot. This score was used because it relies solely on language-independent features and some of the accounts captured do not post in English. The reliability metric used was the IMP score metric. We chose this particular metric to provide a fair comparison, as Botometer scores are more focused on accounts and content and do not use social feedback features [243]. In addition, we also normalize the score since Botometer outputs scores between 0 and 1.

We compute the universal bot score in a large sample of unreliable accounts (n $\approx$ 50000). The percentage of bots identified is approximately 11% since that, according to the authors, accounts are only considered bots if their score is higher than 0.5. Figure 4.3 presents the contrast between the obtained universal score and the IMP metric for each account.

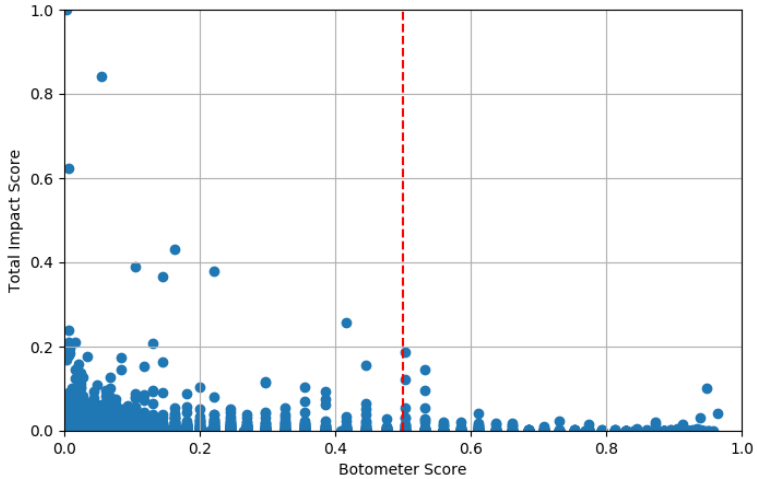


Figure 4.3: Comparison between IMP and Botometer score for the accounts retrieved. The red trace indicates the separation between human-operated and bot accounts according to Botometer. Above that value (Botometer score > 0.5), accounts are considered to be bots.

We also investigate whether there is a significant correlation between the Botometer scores

and the BEH metric values. The hypothesis is that bots have a higher impact on the network (and therefore score higher on our metric) than human-operated accounts.

Since the distribution is not normal, we evaluate the correlation between the IMP total score and the Botometer universal score using the Spearman correlation. The value obtained is 0.191 ($p\text{-}value<0.0001$) indicating a lack of correlation between bot and reliability scores.

To better understand this result, we further examined some of the accounts in the sample. Thus, we manually analyzed the accounts with the highest Botometer score and the normalized BEH metric. Table 4.1 contains some characteristics of the 10 accounts with the highest Botometer score while Table 4.2 refers to the top 10 accounts according to our metric.

Table 4.1: Accounts with the highest Botometer score (B_Score) and their respective reputation score (IMP), number of followers (NFOLLOWERS) and posts (P_Count), and date of creation.

ID	**B_Score**	**IMP**	**NFOLLOWERS**	**P_Count**	**Creation Date**
Bot1	0.965	0.041	305	6.0k	Mar 2017
Bot2	0.958	0.001	175	259	Jan 2019
Bot3	0.952	0.006	3114	11.7k	May 2012
Bot4	0.952	≈ 0	4180	1971	Dec 2016
Bot5	0.948	0.001	253	5.7k	Jun 2017
Bot6	0.948	0.1	27	3	Aug 2014
Bot7	0.944	≈ 0	1275	6	Feb 2019
Bot8	0.944	0.001	44	70	Mar 2014
Bot9	0.942	≈ 0	1	3	Feb 2019
Bot10	0.942	≈ 0	12	1	Dec 2013

Table 4.2: Accounts with the highest reliable score (IMP) and their respective Botometer score (B_Score), number of followers (NFOLLOWERS) and posts (P_Count), and account's creation date. The account in bold was suspended in the analysis process. Thus, the information provided was extracted prior to the analysis and therefore may be outdated.

ID	**IMP**	**B_Score**	**NFOLLOWERS**	**P_Count**	**Creation Date**
Unr	1	0.003	5044	46.8k	Oct 2014
Unr2	0.842	0.054	9751	358.9k	Dec 2018
Unr3	0.622	0.006	863	20.3k	May 2019
Unr4	0.432	0.162	1047	17.9k	Mar 2018
Unr5	0.390	0.104	3706	83.4k	May 2017
Unr6	0.378	0.221	12120	134.8k	Jul 2016
Unr7	0.365	0.146	6615	20.7k	Nov 2018
Unr8	***0.258***	***0.416***	***22.6k***	***86.2k***	***Jan 2019***
Unr9	0.234	0.006	3967	321.8k	Jun 2013
Unr10	0.210	0.016	843	13.1k	Jun 2018

There are several observations that are important to highlight from Table 4.1 and 4.2. First, the bottom half of the top accounts with the highest bot score have a lower than average number of posts ($\leq$70) and a low number of followers (with the exception of Bot7). Furthermore, some accounts are classified as bots with high confidence but the number of posts and followers is below 10. This means that even if these accounts are labeled as bots, their current impact on the Twitter ecosystem as disseminators of unreliable content is low. Shifting our analysis towards Table 4.2 we can see that none of the unreliable accounts with the highest scores were classified as bots (bot score $>$ 0.5) by Botometer. This fact strengthens the importance of these metrics, as accounts that are often sorted out as bots can still pose a threat to the Twitter ecosystem. By manually inspecting the accounts, we can conclude that Unr10 and Unr9 are posting false information and exhibiting human-operated behavior, as they contain original tweets, original profile images (not found on other websites by using reverse image search), and personal information. Un2, Unr4, Unr6 and Unr7 also exhibit human behavior although they do not have an original profile picture. Unr5 is a non-English account whose unreliable tweets captured are English retweets. Finally, accounts Unr1 and Unr3 are the exception, as they are social media accounts for two of the websites included in OpenSources (and thus they achieve a high unreliable score). Two important conclusions can be derived from this analysis. The first is that current bot detection systems are not sufficient to detect unreliable accounts mainly because they were originally designed to detect accounts that are operated automatically. Similarly, some of the top accounts that are classified as bots do not have a significant impact on the network due to their low number of connections and publications. On the other hand, Table 4.2 highlights the importance of metrics and systems to detect unreliable accounts, as they are often not detected by bot/spam detection systems. Furthermore, human-operated unreliable accounts represent a large portion of the unreliable accounts analyzed ($\approx$ 89%), which reinforces that bot detection systems are not sufficient to prevent the detection of unreliable accounts.

4.1.1.3 Twitter Unreliable Accounts Evaluation

The final case study examines the type of content captured in Twitter using their Search [235] and Filter API [238]. While the former allows extracting popular tweets based on certain keywords, the latter allows retrieving posts in real-time (consequently increasing the volume of data). The main goals of this case study are to demonstrate the capability of our methodology on capturing accounts that are disseminating unreliable content and to investigate the current state of Twitter with respect to unreliable accounts that are still active and whose posts are still relevant and returned by Twitter. Due to the time period in which the experiment was conducted, only a sample of knowledge (about the accounts acquired from our methodology) was used.

Data Retrieval Although Twitter is actively fighting unreliable content and the corresponding accounts that distribute it, it is still important to analyze whether such unreliable content is present in the posts retrieved by their APIs. This is due to the implications that the data retrieved from the API can have on real-world. For example, the data collected in the Filter API is commonly used for data analysis of real-time events [9, 195, 136] while the results from the Search API are similar to the results users get when they use the search bar on the Twitter platform [235].

For this analysis, we created 4 different datasets, combining two different extraction methods with the two API endpoints previously mentioned.

The first extraction was based on the trending topics provided by Twitter [236]. During 12 days (from November 29 to December 11, 2018), the daily trending topics were extracted. However, since the Search API has two query types (recent and relevant) a maximum of 200 tweets were extracted for each. Regarding the Stream API, we opened a 60-second capture window for each trending topic on a daily basis. This process yields the first and second datasets.

The second retrieval method supports the hypothesis that the more relevant, newsworthy and relevant the topics, the more publications with unreliable content (and consequently the accounts that disseminate it) are returned. The third and fourth datasets were extracted using manually selected topics, taking into account the controversy and journalistic relevance at the time. The keywords used were the following: "trump", "bolsonaro", "vaccines", "syria", and "lgbt". The topics are mainly related to the following events: the sworn of controversial Brazil President Jair Bolsonaro [221], the false information linking vaccines to autism [182], the Syrian War and the decision by President Donald Trump to pull out the United States troops from the conflict [19], and the persecution of the LGBT community in several countries [8, 110].

The extraction procedure for both APIs follows the same methodology as for the first dataset. However, the start and end dates are different (January 4 - January 16, 2019). Nevertheless, we only computed the account metrics with data knowledge extracted before this time interval.

Results In this section, we present the results of analyzing the data extracted based on the previous knowledge acquired. Thus, we showcase the accounts that are already presented in the database with the respective metrics.

The different datasets are named according to their retrieval process. Thus, regarding the stream API, there are the TWST_TOPICS for the trending topics and TWST_SELECTED for the selected keywords dataset. Similarly, the datasets for the search API are TWSRCH_TOPICS and TWSRCH_SELECTED.

Table 4.3: Results for the average score of the Post Based metric ($PCOUNT$) for each different category in each dataset. Results in bold shows the dataset with highest value for each category.

PCOUNT	TWSRCH_TOPICS	TWSRCH_SELECTED	TWST_TOPICS	TWST_SELECTED
Bias	4.98	**10.27**	6.27	8.59
Clickbait	**5.61**	2.63	2.35	2.51
Fake	1.19	**2.76**	1.43	2.18
Hate	0.20	**0.48**	0.30	0.40
JunkSci	0.38	**5.00**	0.54	0.79
Unreliable	1.65	**3.84**	2.20	2.39
ALL	14.03	**24.98**	13.09	16.85

The first analysis conducted was regarding the percentage of unreliable accounts (present in the database) found in the datasets. In this analysis, an account that posted at least one tweet from one of the different selected OpenSources categories is considered unreliable (i.e. $PCOUNT \geq 1$). The TWSRCH_TOPICS contains 543 tweets from accounts already identified in our database. This represents about 4% of the extracted dataset (total size is 12455). In the TWST_TOPICS, the number of unreliable accounts detected is slightly higher corresponding to approximately 6% of the total dataset.

The datasets built with selective keywords reach the highest percentage of accounts that disseminate unreliable content. In the TWSRCH_SELECTED, 25% of the accounts were already present in our database. This corresponds to a total of 2180 tweets out of 8417. Finally, regarding the TWST_SELECTED, 222337 tweets were captured concerning accounts already flagged. This corresponds to a total of approximately 36% of the dataset and it is the highest percentage of all datasets analyzed.

We apply the developed metrics to the accounts discovered in each dataset to further investigate the impact they had on the sample of tweets retrieved. We use the average of the accounts' scores for each of the different categories. The results are presented in Table 4.3, 4.4 and 4.5 for the post-based, the behavior / behavior with social feedback and the impact/impact with social feedback, respectively.

In all the metrics analyzed, the impact of unreliable accounts is higher when we use controversial and journalistic relevant keywords.

Regarding the individual types of unreliable content, the "bias" class outperforms in all scores. This is also to be expected, since the number of tweets in this category is very high. However, it also means that there is significant exposure to extreme biased content in the Twitter ecosystem. As a matter of fact, in the datasets where the selected keywords were used (which corresponds to the tweets that appear in users' feeds) the dominance of biased content is the highest, with unreliable accounts posting an average of 10 publications related

Table 4.4: Results for the average score of the Behavior (BEH) and Behavior with Social Feedback (BEH_SF) metric for each different category in each dataset. Results in bold shows the dataset with highest value for each category for each individual metric.

	TWSRCH_TOPICS		TWSRCH_SELECTED		TWST_TOPICS		TWST_SELECTED	
	BEH	BEH_SF	BEH	BEH_SF	BEH	BEH_SF	BEH	BEH_SF
Bias	0.032	0.144	**0.082**	**0.317**	0.049	0.192	0.062	0.258
Clickbait	**0.018**	0.054	0.014	0.061	0.014	0.062	0.015	**0.065**
Fake	0.008	0.033	**0.018**	0.063	0.012	0.052	**0.018**	**0.064**
Hate	0.002	0.007	**0.005**	0.018	0.003	0.014	0.004	**0.019**
JunkSci	0.004	0.014	**0.030**	**0.091**	0.005	0.017	0.006	0.020
Unreliable	0.009	0.037	**0.034**	**0.116**	0.013	0.054	0.017	0.070
ALL	0.073	0.288	**0.182**	**0.666**	0.098	0.390	0.123	0.495

Table 4.5: Results for the average score of the Impact (IMP) and Impact with Social Feedback (IMP_SF) metric for each different category in each dataset. Results in bold shows the dataset withh highest value for each category for each individual metric.

	TWSRCH_TOPICS		TWSRCH_SELECTED		TWST_TOPICS		TWST_SELECTED	
	IMP	IMP_SF	IMP	IMP_SF	IMP	IMP_SF	IMP	IMP_SF
Bias	0.152	0.675	**0.364**	**1.433**	0.210	0.863	0.275	1.143
Clickbait	0.085	0.241	0.064	0.273	0.064	0.275	**0.065**	**0.283**
Fake	0.034	0.151	0.080	0.283	0.058	0.243	**0.083**	**0.299**
Hate	0.008	0.033	**0.021**	0.076	0.014	0.061	0.019	**0.083**
JunkSci	0.020	0.067	**0.125**	**0.372**	0.020	0.072	0.026	0.087
Unreliable	0.039	0.164	**0.143**	**0.496**	0.058	0.241	0.075	0.305
ALL	0.338	1.33	**0.797**	**2.933**	0.424	1.755	0.544	2.199

to this category.

When considering the age of the account and the tweet, the score drops significantly. This means that, on average, the accounts included in the various datasets are not new users, but rather established users in the Twitter ecosystem. For example, for the TWST_SELECTED the more frequent age values of unreliable accounts are in the range of 20 to 30 months, followed by 120 to 140 months. The range of frequent values is similar for the other datasets analyzed.

When we add social feedback to the behavior metric, the values triples in each class. Hence, although the tweets captured in the database were "recent", their social feedback highly affects the accounts retrieved, demonstrating the rapid spread of unreliable tweets.

The main proposed metrics are assessed in Table 4.5. Although there are no major changes in the predominance of scores on the different datasets compared to Table 4.4, it is important to highlight the differential growth of the metrics with the addition of INFLUENCE. The increase ranges from 0.3 to 0.6 from BEH to IMP and 1.0 to 2.3 in their social feedback versions. Because the more frequent influence scores in all datasets range from 2.5 to 5, there are no significant differences between BEH and IMP metrics. Thus, the use of these two metrics for analyzing a set of accounts in an aggregated fashion is redundant. However, for the analysis of individual accounts, these two metrics can provide useful insights to measure the impact and influence of these accounts in the network.

4.1.2 Conclusions

In this section, we have described a set of knowledge-based metrics to measure the behavior of accounts and their impact on the network in terms of their reliable or unreliable behavior. These are aimed towards the development of an unreliable account detection system where information about a given account can be summarized based on its past behavior and its current impact on the network. To better illustrate the usefulness of these metrics and their necessity in a context where human- and automatic-operated accounts distribute unreliable content, two case studies where conducted: analyzing the accounts with the highest scores on one of the proposed metrics and comparing a sample of accounts with a bot detection system. Based on these case studies, we can conclude that 1) the proposed methods and metrics are able to capture and classify unreliable accounts that may pose a threat to the Twitter ecosystem, and 2) the metrics are useful because they can identify unreliable accounts that are not captured by bot detection systems such as Botometer.

Finally, the third case study aims to prove the need for such a system by showing the prevalence of unreliable accounts in Twitter. The results show that unreliable accounts are responsible for 4% and 6% of the tweets referring to trending topics and extracted

using Search and Stream APIs, respectively. However, when journalistically relevant and controversial keywords are used, this percentage increases to 25% and 36%. When evaluating the type of unreliable accounts using our main metrics (IMP and IMP_{sf}), the Search API with journalistically relevant and controversial keywords has the highest values. This means that while there is a higher percentage of unreliable tweets in the TWST_SELECTED, when considering other factors such as the number of unreliable posts shared, social feedback, and influence, it is in the TWSRCH_SELECTED that these metrics reach their highest value.

However, if we consider a more pragmatic scenario in which the data and metrics presented in this chapter are applied in a real-world detection system, its accuracy is as good as the knowledge extracted. Therefore, due to the ever-growing number of accounts in Twitter, limited storage, computational power, and API rate limits, this knowledge will only ever be a sample of the total information available in the social network. Therefore, when developing a detection system, one should not limit itself to this fraction of knowledge. Instead, it must be complemented by the ability to predict whether an account is likely to be reliable or unreliable based on its attributes and content at the time.

Therefore, in the next section, we explore the problem from a machine learning perspective and propose a model for detecting unreliable accounts based on the available account information and content.

4.2 Classification-Based Approach

In this section, we address the problem of automatically labeling a social media account as reliable or unreliable. Unlike other studies referenced in Chapter 2, we focus on the detection of reliable and unreliable accounts regardless of how they are operated.

In this scenario, and working towards the implementation of a fully functional unreliable accounts detection system, it is important to consider that the content available for analysis may not always be the same across all social media accounts. For example, recently created human-operated accounts (unreliable or not) may have posted a lower number of tweets than older and more established accounts. This factor is often ignored in bot detection studies, as the number of posts used to train and test the models is usually not mentioned.

Therefore, we hypothesize that variation in the number of posts may negatively affect model performance and that the assumption that this variation occurs in a more realistic scenario is valid. Consequently, although a classification model using content features can be trained and tested with the same volume of tweets in all accounts, in a more pragmatic scenario it must be expected that this volume of publications will vary. Hence, if a model is trained with data from accounts with a large volume of tweets, its prediction performance may degrade for accounts with a lower volume of posts.

Considering this variation in the volume of tweets and the aforementioned problem, another important aspect arises: due to the difference in publishing frequency between bot and human accounts, a fixed number of tweets may correspond to a larger time interval for a human account than for bot account, since the latter publishes at a higher frequency. In fact, one of the most commonly used features in bot detection studies is the frequency of tweets or average tweets per day, as the values of human and bot accounts differ on this metric. Also, in a previously published work on unreliable accounts (humans and bots) [93, 94] we also found that top unreliable accounts have very different posting frequencies for the last 3200 tweets (which is the maximum number of tweets allowed in the standard Twitter API).

Therefore, for the aforementioned problem, it is not only important to study the performance of the models by restricting tweets by volume, but also by time interval. Consequently, in this work, we evaluate how the performance of the models is affected by the variation of the content, based on two types of batches: the volume (number of tweets) and the time (time intervals in days). The main goal is to identify the best conditions to maximize performance on the classification of unreliable accounts regardless of the amount/volume of data provided.

Our proposed solution evaluates each model individually with different volume and time interval batches, and compares these approaches with ensemble models that adapt according to the data available for each account. In the next section, we explain the experimental setup and describe in more detail the experiment conducted.

4.2.1 Experimental Setup

We have developed an experimental setup (illustrated in Figure 4.4) to address the problem described in the previous section and to lay the groundwork for our proposed solution. Each individual component is further detailed in the next sub-sections.

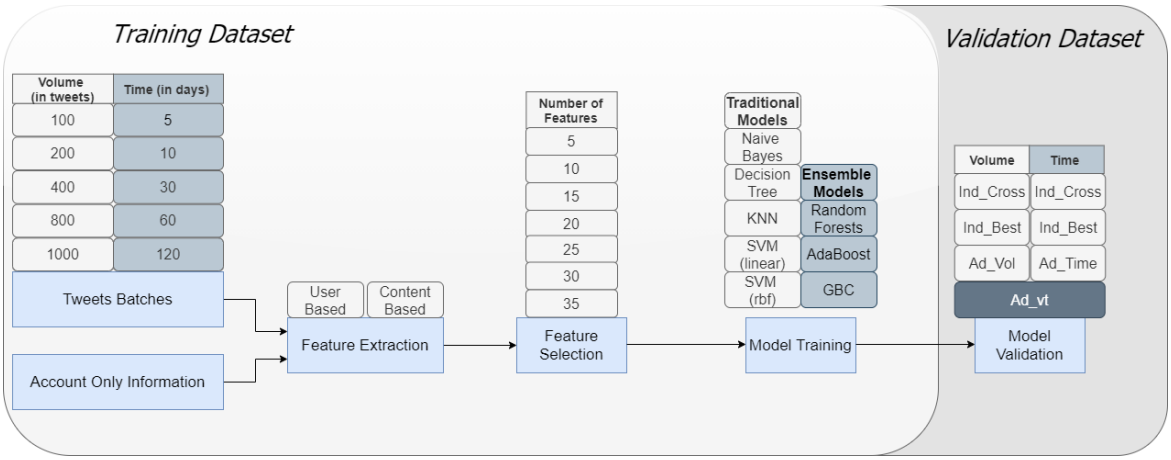


Figure 4.4: Diagram of the experimental setup

4.2.1.1 Data Extraction

The data used in the experiments presented in this chapter were extracted according to the methodology described in Section 3.1.1. The tweets used to identify reliable and unreliable accounts were retrieved from July 2019 to April 2020. These tweets were then aggregated by account according to the methodology described previously. However, instead of using the scores as a prediction label, we chose to approach the problem as a binary classification task. The main reason for this decision were the limitations in data extraction and score assignment. More specifically, since we do not have access to a full coverage of public tweets and we limited ourselves to the websites in OpenSources and MediaBiasFactCheck, it is not guaranteed that a given account is disseminating more unreliable or reliable content than the score indicates (i.e. has a higher score than the one captured). In other words, due to these constraints, we can only ascertain their unreliable/reliable class since we can determine with some degree of confidence that each account has disseminated at least x posts from the class to which it was assigned. Consequently, the label assignment for each account (a) was made according to the following equation.

$$\text{label}_a \begin{cases} unreliable \iff score \leq -10 \\ reliable \iff score \geq 10 \end{cases} \tag{4.7}$$

We defined the value to be greater than or equal to 10 by experimentation. On one hand,

our goal was to maximize the range of scores for both reliable and unreliable accounts (since restricting the data to accounts with a high score would likely result in a dataset consisting of a majority of reliable and unreliable bots). On the other hand, relaxing the score criteria to values in the interval $[-10, 10]$ would lead to the inclusion of accounts that are majority neutral and have been captured only a small number of times during the extraction process.

For each account, we then proceed to crawl each timeline and extract the maximum number of tweets that the API allows (approximately 3200 per account).

A balanced sample of 1000 reliable, and unreliable, accounts was extracted. The data was then randomly divided into a training and a validation datasets, with a 70-30 percent split.

4.2.1.2 Topic and Keyword Analysis

To better understand the extracted data and to ensure that the models are not affected by the diversity of topics in each class, we perform a topic and keyword analysis. The main motivation for this analysis is the hypothesis that, if the topics being discussed by reliable and unreliable accounts were completely different, then the trained models could accurately distinguish the accounts based on the topics discussed rather than on their unreliable/reliable behavior.

We began by performing a simple topic analysis using LDA [26]. We applied standard cleaning techniques to the tweets, but chose to retain the hashtags and mentions since hashtags can be used as topic aggregators and mentions can refer to important entities such as *@realdonaldtrump*. We aggregated all tweets belonging to accounts in each class and extracted 10 topics. The results are presented in Table 4.6.

When analyzing the keywords for each topic, it is clear that the discussion for both reliable and unreliable accounts revolves around the new coronavirus pandemic (Covid-19) and former United States President Donald Trump. We consider these results predictable because Covid-19 has been a relevant topic in the news since 2019 [159] and Donald Trump has also been a subject of journalistic attention since the beginning of his presidential campaign in 2015 [83]. In addition, both topics were dominant in the 2019 news landscape due to the United States presidential election and President Trump's administration's response to the pandemic [152].

To complement our topic analysis, we used the YAKE system [37] to extract relevant keywords from individual tweets and aggregate them in terms of reliable and unreliable data. YAKE assigns a score to each keyword representing its importance in the text. Thus, for each account, we created a lexicon with all the extracted unigram keywords, and added the scores of the repeated keywords. We then represent the top 10 keywords for each account. Next, a similar procedure was applied to determine the most relevant keywords for each class.

Table 4.6: Topics and associated words for the totality of tweets extracted from unreliable and reliable accounts. Associated words in bold represent the ones that are included in some topic in both reliable and unreliable classes

Reliable		**Unreliable**	
Topic	Associated words	Topic	Associated Words
0	**"trump"**,**"president"**,**"coronavirus"**,"todays", "features","mentions","homepage",**"realdonaldtrump"**, "trumpindex",**"donald"**'	0	**"trump"**,**"president"**,**"realdonaldtrump"**,"biden", **"coronavirus"**,"thank","pelosi",**"donald"**, "working" ,"bill"'
1	"newspicks" ,"business",**"coronavirus"**,"daily", "thanks","real","market" ,"social" , "estate","work"	1	**"coronavirus"**,**"realdonaldtrump"**,**"trump"**,"great", **"president"**,"stop","whitehouse","change", right",**"going"**
2	"times" ,**"israel"**,**"coronavirus"**,**"covid"**, **"trump"**,"india" ,"google" ,**"going"**, "greece" ,"tribune"'	2	**"good"**,**"time"**,"youtube",**"please"** , "look",**"coronavirus"**,**"first"**,"everyone", "many",**"make"**
3	**"make"** ,**"coronavirus"**,**"know"** ,"help" , **"need"** ,"para" ,"adventure" ,"time", **"covid"**,"book"'	3	**"know"** ,**"china"** ,**"trump"**,**"president"**, **"coronavirus"**,"obama","virus" ,"charliekirk", **"realdonaldtrump"**,"could"
4	"time" ,"best","books" ,**"last"** , "year" ,**"make"** ,**"first"** ,"week", "really","music"	4	**"going"**,"home","still",**"would"**, **"realdonaldtrump"**,**"love"** ,"think","stay", "could","time"
5	**"coronavirus"** ,"cases" ,**"covid"**,**"need"** , **"take"** ,**"please"**,**"love"** ,"much" , **"would"** ,"pakistan"	5	**"realdonaldtrump"**,**"trump"**,"democrats",**"coronavirus"** , "nothing","america",**"need"**,"maga" , "time","qanon"'
6	"business" ,"antgrasso" ,"covid" ,"real", "market" ,"dubai","global" ,"digital" , "health" ,"middleeast"'	6	"money",**"house"**,"white",**"realdonaldtrump"**, **"trump"** ,"give","congress",**"coronavirus"**, "bill",**"israel"**
7	**"coronavirus"** ,**"realdonaldtrump"**,**"house"** ,**"know"** , **"first"** ,**"china"**,**"good"** ,**"time"** , "morning" ,"states"	7	**"trump"** ,"never","bernie",**"would"** , **"realdonaldtrump"**,"biden" ,"true" ,"sanders", **"covid"**,"president"'
8	**"trump"**,"times","york" ,"banking" , "openbanking" ,**"coronavirus"** ,"transit" ,"open" , "transportation","todays"	8	**"trump"** ,**"coronavirus"**,**"realdonaldtrump"**,**"would"**, **"covid"**,**"make"** ,**"last"** ,"think", **"know"**,**"take"**
9	**"coronavirus"**,**"trump"**,**"covid"**,"times" , "daily","todays" ,"cruise","features", "breaking","trumpindex"'	9	**"coronavirus"**,"virus" ,"chinese" ,**"covid"**, "realjameswoods","american",**"realdonaldtrump"**,**"trump"**, **"president"**,**"would"**

More specifically, the keywords for all accounts were added to a new lexicon, representing a list of pairs (keyword, score), again, summing the scores of the repeated keywords.

Figure 4.5a and 4.5b present the wordclouds in terms of the reliable and unreliable keywords, where the size is represented by the aggregated score of each keyword.

Similar to the topic analysis, keywords such as "trump" and "coronavirus" achieve high relevance in both reliable and unreliable tweets. In addition, terms such as "bernie", "democrats" and "biden" also appear in both wordclouds and are related to the most important topics discussed.

Both topic and keyword analysis demonstrate the similarity of topics discussed by reliable and unreliable accounts. Consequently, models trained and tested on these data will not be influenced by topic diversity between classes (i.e. the content of tweets from reliable accounts being significantly different from unreliable accounts).

(a) Reliable wordcloud *(b)* Unreliable wordcloud

Figure 4.5: Wordclouds for the most relevant keywords in each class. Keywords were extracted and ranked using YAKE [37]. The size of each keyword represents their relevance weight where the larger the font size, the more relevant the keyword is in that specific class.

4.2.2 Tweets Batch Size and Time Intervals

To conduct experiments that accurately classify unreliable and reliable accounts, we created batches of tweets based on volume (number of tweets), and time intervals (days).

The volume batches were ordered from the most recent to the oldest tweet. For this experiment, we selected 5 different volume batches: 100, 200, 400, 800, and 1000. We highlight that the lower batches are subsets of the higher ones, meaning that $\text{batch}_i \subseteq \text{batch}_{i+1}$ where $i \in \{100, 200, 400, 800\}$.

Although the majority of accounts contain at least some posts/tweets, we also consider the scenario where no tweets are present (i.e. no tweet information is used and only account-related features are extracted). For example, for newly created accounts or accounts whose tweets have been deleted due to the measures imposed by Twitter to remove publications with misinformation [108].

For the time-based batches, we ordered the tweets from the most recent to the oldest and extracted different batches based on 5, 10, 30, 60, and 120 day intervals from the most recent post.

4.2.3 Feature Extraction

The features extracted can be aggregated into two groups: account- and content-based features (concerning each batch).

4.2.3.1 Account-based features

These features only use information about the accounts, without any knowledge of the type of content being published or shared. In addition to features commonly used in bot detection systems, such as the number of followers and friends, verification status, age, and presence of default profile, we also used the number of each type of Twitter-entity represented in the account information fields (such as URLs, hashtags and mentions), the sentiment of the account description, and whether the description fits into any political domain (using a political lexicon from Oxford Topic Dictionaries [4]. With respect to the account description, we also extracted the number of emojis and emoticons, as well as the percentage of capital letters and entities (i.e. persons, organizations, and locations) identified by NLTK [25]. We also extract a readability score from the description to be used as a feature, and a large number of lexical categorization features using the Empath tool [73]. Furthermore, we used Google pre-trained Word2Vec model [150] to extract embedding vectors based on the accounts' description text. Only words that are included in the vocabulary of the Word2Vec model are considered.

4.2.3.2 Content-based features

The content-based features are derived based on information from a set of tweets. We start by reusing some of the features used in the account description and apply them to tweets (entities, emoticons, emojis, percentage of capital letters). In addition, we quantify the different parts of speech and emotions associated with the post (using the NRC emotion lexicon [155]), count the number of URLs, hashtags, and mentions, and the average word length per tweet. We use the Word2Vec model to extract 300-word vectors for each tweet. These features were computed individually for each tweet and then the average was calculated for the entire batch/time interval. Similar to the account features, only words that were included in the Word2Vec vocabulary were considered. Also included as a feature: the frequency of posts and the retweet ratio within each group of tweets.

We extracted these features for each batch and time interval. Therefore, we created 11 different datasets (5 volume-based + 5 time-based + 1 for zero tweets).

4.2.4 Feature Selection

For each of the generated datasets, we applied a feature selection process to determine the best n features. We measured the correlations between each pair of features and removed the highly correlated ones. We set a threshold of 0.95 for this procedure. Finally, we calculated

[4] `https://www.oxfordlearnersdictionaries.com/topic/politics`

the best features using the mutual information score. The features were sorted by descending order, and the most relevant ones were used. In this work, we opted to experiment with different feature sizes. We chose seven different sets: the top 5, 10, 15, 20, 25, 30, and 35, where, similar to the batches, smaller sets of features are subsets of larger ones. Combined with the different batch sizes and time intervals, we have 77 different training datasets of content features.

4.2.5 Model Training and Evaluation

The models used for training can be divided into two different categories: individual classification models and ensemble models. Our selection process took into account the current state of the art in bot detection (see Chapter 2). Therefore, the following classification models were selected: SVMs (with linear and radial basis function kernels), Naive Bayes, Decision Trees, KNN, Random Forest, Gradient Boosting, and AdaBoost. Several of these models achieved good performance in bot detection studies [61, 1, 179].

For the performance evaluation of each model, we used a weighted F1-score where the weight is defined by each class support. The weighted F_1 score is presented in Equation 4.8 where F_{1_u} and F_{1_r} are the F1-scores for the unreliable and reliable accounts (respectively) and $|u|$ and $|r|$ the support (i.e. number of entries) in each class.

$$F_{1_{weighted}} = \frac{(F_{1_u} \times |u|) + (F_{1_r} \times |r|)}{|u| + |r|} \tag{4.8}$$

Each model is evaluated considering different batches of tweets and a different number of features. The evaluation consists of two different stages. In the first, for each batch and number of features, a model is built and evaluated using 5-fold cross-validation in the same settings used in the training. In the second stage, we evaluate the models in a cross-batch scenario where a model trained in a particular batch is evaluated using the data from the remaining batches.

4.2.6 Validation

Since our goal is to evaluate how the performance of traditional approaches compare with adaptive solutions that consider the best model based on the volume and time, we used the results of the previous evaluation to define 7 different models for the validation step.

- Ind_best_volume and Ind_best_time: the best individual models regardless of batch or features. In other words, the models with the highest weighted F1-score in the first stage of the evaluation (models trained and evaluated in the same batch). Two models

are created based on the volume batch and the time batch, respectively. These models can be considered as baseline because their training and evaluation are similar to those in several studies described in Chapter 2.

- Ind_cross_volume and Ind_cross_time. These are the two models that perform the best in a cross-batch scenario. More specifically, the model where the average of the weighted F1-score of all batches is the highest. Similar to the first case, there is one model for each type of batch.

- Ad_volume and Ad_time: Ensemble of models that adapt to the number of tweets from each account. Depending on the number of tweets, the model with the highest performance for that batch is selected independently of the number of features. Ad_volume adapts according to the volume of tweets while Ad_time adapts with respect to the interval of days of the set of tweets given.

- Ad_vt: This solution combines the previous two and uses only one model – the one that achieves the highest performance on the evaluation step. In other words, for an account a with a number of tweets n_a, this model computes the interval time in tweets n_a and selects the best time model for this interval. The same is done for the volume (i.e. the best model is selected for the volume n_a). Finally, only the model that scored the highest weighted F1-score is used for classification.

The evaluation of the best models is performed in the validation dataset, which refers to accounts that were not used in the previous evaluation. The performance of these models is validated considering a pragmatic social network scenario where the number of posts from accounts varies. Thus, we assign a random number of tweets from the defined batches to each account in the validation dataset.

Moreover, to improve the robustness of the experiment conducted, the results presented are the average of 5 iterations through the test set, with a new randomly generated volume calculated for each account at each iteration. In addition, we considered 3 different weighted probabilities for the generated volume:

- High volume distribution: 25% probability of the volume of the account is in the interval [100,200] and 75% probability is in [400,800,1000]

- Equal volume distribution: equal probability for each batch

- Low volume distribution: 75% probability of the volume of the account is in the interval [100,200] and 25% probability is in [400,800,1000]

In the specific case where accounts have no tweets, a fair comparison with the other models is not possible. In addition, due to the limited size of the validation dataset, we chose to

validate the case where no content is available separately. This allows us to highlight the performance of the best model on never-before-seen data.

4.2.7 Results

In this section, we present the results of applying the workflow presented in Figure 4.4. We start by analyzing the features selected in each defined set for each different batch. We then evaluate the selected models in the two scenarios designed in the workflow (individual batches and cross-batches). Finally, we present the performance of the individual, cross-batch, and adapted models using the validation data.

4.2.7.1 Feature Analysis

We start by analyzing which features are selected when considering the absence of tweets (i.e. account features only). The dataset of 5 features includes the friend-to-followers ratio (users_ratioFF) and the average of "favorites" and tweet frequency (users_favouritesPerDay and users_tweetsperday). These features are often used in bot detection systems. The other 3 features are word vectors from the pre-trained Word2Vec model applied to the description of the account. These types of features are less common in bot-detection approaches since some works focus only on the presence of a description text [55] or some particular characteristics, such as text length [243, 67]. When considering a larger number of features, those derived from word vectors are predominant, with the exception of the neutral and compound sentiment score in the account description (users_sentimentDescription_neutral and users_sentimentDescription_compound). By analyzing the selected features, when no content features are used, we can observe the importance of the description-based features. However, it is important to point out that features commonly used in the bot detection task, such as the number of tweets per day, and the ratio of friends to followers, also play an important role even when only a small set of features is considered.

When analyzing the volume-type batches, the inclusion of tweet-derived features shifted the selection of the most important features significantly. In fact, regardless of the volume batch and number of features considered, all features are derived from the Google pre-trained Word2Vec model. It is important to emphasize that the selected vectors change when considering different volume batches, suggesting that the number of tweets on each account has an impact on the selected word vectors.

When considering the time-based batches, the same pattern emerges. In other words, the relevance of the word vectors is prevalent across the different batches and number of features. However, one account indicator that eventually emerges as the number of features increases is the number of favorites per day. In the 5-days batch, it is selected among the top 10

features. However, in the 10- and 30-days batches it appears only among the top 15, and in the 60-days batch it appears only among the top 25 features. Finally, in the 120-days batch, it appears among the top 20 features. It is difficult to explain this phenomenon, but we assume that a large number of days allows the presence of multiple topics discussed in these tweets. Hence, with a larger lexicon and topics being changed, it is likely that some word vectors do not perform as well, and therefore give way to other types of features.

In summary, general features used in bot detection tasks are important when there is no content available (i.e. published tweets), as they remain among the top 5 of most important features. However, when account content is considered, the importance of word embedding features outweighs all others, regardless of the type of batch and volume of information considered. Hence, in the absence of tweets, common features derived from the bot detection task are important, but are overshadowed by the word-embedding features when content is considered. A summary of the selected features and their respective importance is presented in Table 4.7.

4.2.7.2 Model Evaluation

We proceed to evaluate the different models proposed. The objective of this section is twofold. First, we evaluate the change in performance that each model exhibits with the addition of features and batches. Then, we evaluate how the best models, trained in one particular batch, perform when tested in the remaining ones.

Individual Batches We begin by evaluating the models in individual batches (i.e. models are trained and tested using the same batch). To establish a baseline performance, and since the training data is approximately balanced, we use a dummy model that randomly assigns a classification label. This baseline model achieves a weighted F1-score of 0.53.

The evaluation of the different models is presented in Figure 4.6 for the case where accounts' tweets are not included, and in Figures 4.7 and 4.8 for the volume-based and time-based batches experiments, respectively.

Regarding the scenario where only account features are used, Random Forests and Gradient Boost Classifier (GBC) are the models that, on average, perform the best. The model with the worst performance is Naive-Bayes. This model's performance increases with the addition of more features but remains below the baseline in all cases except when the number of features is 35. The best performance is obtained with 30 features (0.71), although there is only an increase of 0.2 over the best model with 15 features and of 0.4 over the best model with 10 features.

As for the different models trained in the volume batches, Random Forests show more stable

Table 4.7: Feature importance of the 35 top features for each batch. For the sake of comprehension, the feature importance of word embedding features is presented in terms of average and standard deviation.

Account Only

Batch	**Feature Names**	**Feature Importance**	
0 tweets	users_ratioFF	0.0652	
	users_favouritesPerDay	0.0548	
	users_sentimentDescription_neutral	0.0486	
	users_tweetsperday	0.0418	
	users_sentimentDescription_compound	0.0378	
	Word Embedding Features	**Average**	**Standard Deviation**
	user_description (30 vectors)	0.044082	0.008054

Volume - Based

Batch	**Feature Name**	**Feature Importance**	
	Word Embedding Features	**Average**	**Standard Deviation**
100 tweets	tweets_w2v (35 vectors)	0,0996	0,0138
200 tweets	tweets_w2v (35 vectors)	0,1054	0,0174
400 tweets	tweets_w2v (35 vectors)	0,1094	0,0186
800 tweets	tweets_w2v (35 vectors)	0,1139	0,0193
1000 tweets	tweets_w2v (35 vectors)	0,1114	0,0148

Time - Based

Batch	**Feature Name**	**Feature Importance**	
5 days	users_favouritesPerDay	0,1103	
	Word Embedding Features	**Average**	**Standard Deviation**
	tweets_w2v (34 vectors)	0,1047	0,0181
10 days	users_favouritesPerDay	0,1102	
	Word Embedding Features	**Average**	**Standard Deviation**
	tweets_w2v (34 vectors)	0,1098	0,0158
30 days	users_favouritesPerDay	0,1102	
	Word Embedding Features	**Average**	**Standard Deviation**
	tweets_w2v (34 vectors)	0,1156	0,0195
60 days	users_favouritesPerDay	0,1112	
	Word Embedding Features	**Average**	**Standard Deviation**
	tweets_w2v (34 vectors)	0,1187	0,0150
120 days	users_favouritesPerDay	0,1118	
	Word Embedding Features	**Average**	**Standard Deviation**
	tweets_w2v (34 vectors)	0,1200	0,0174

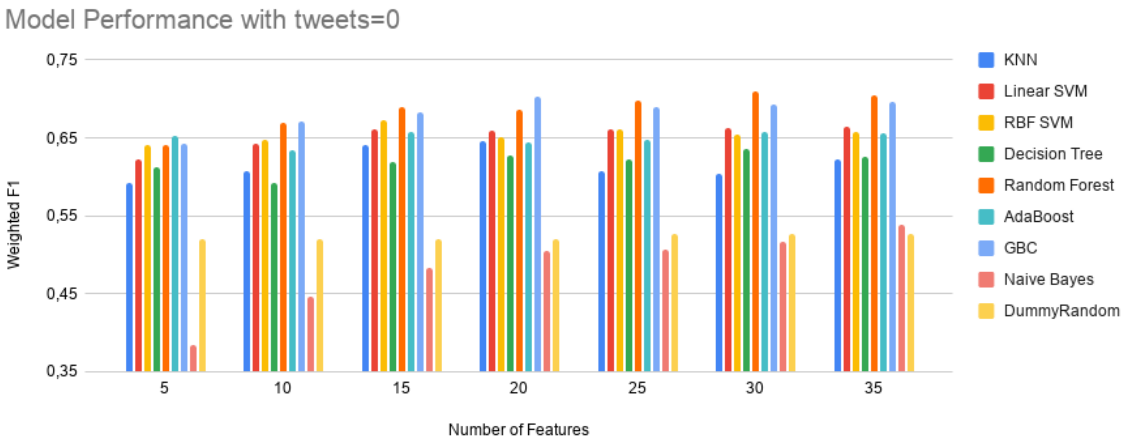


Figure 4.6: Performance variation for the different models and number of features using account-only data (tweets=0).

results through the different settings (number of tweets and features). AdaBoost and GBC also achieve the highest performance in some settings. However, they show larger fluctuations when transitioning from batch to batch. It is also noticeable that for the most of the feature sets considered, increasing the number of tweets does not necessarily lead to equal or better performance of the models. This is visible in the 10, 20, 25 and 35 feature sets where the best models reach their maximum performance in batches of size < 1000.

Moving our analysis to Figure 4.8 which refers to the evaluation of the models in time-based batches, Random Forests also show robust performance in the different cases considered, but it is the GBC model that achieves the highest performance (frequently, in the 30-day batch) in the majority of the different set of features considered (with the exception of the 10-feature set). KNN also shows good results in some batches and feature sets (for example in the 60-day batch with 15 features and 120-days batch with 25 features) although, similar to AdaBoost in the previous experiment, it shows a larger variation in performance between different batches.

The more interesting conclusion from this experiment is that the models tend to reach their highest performance using a 30-day time window, which suggests that more data does not necessary imply an improvement in the models performance. In all the 15, 20, 25, and 30 feature datasets, the performances reach their peaks in this batch. In some cases, this result is not reached again until the models use tweet information from a 120-day time window.

As previously mentioned, we selected the best models for each batch (both time and volume) and performed a cross-batch evaluation. The results are presented in the next subsection.

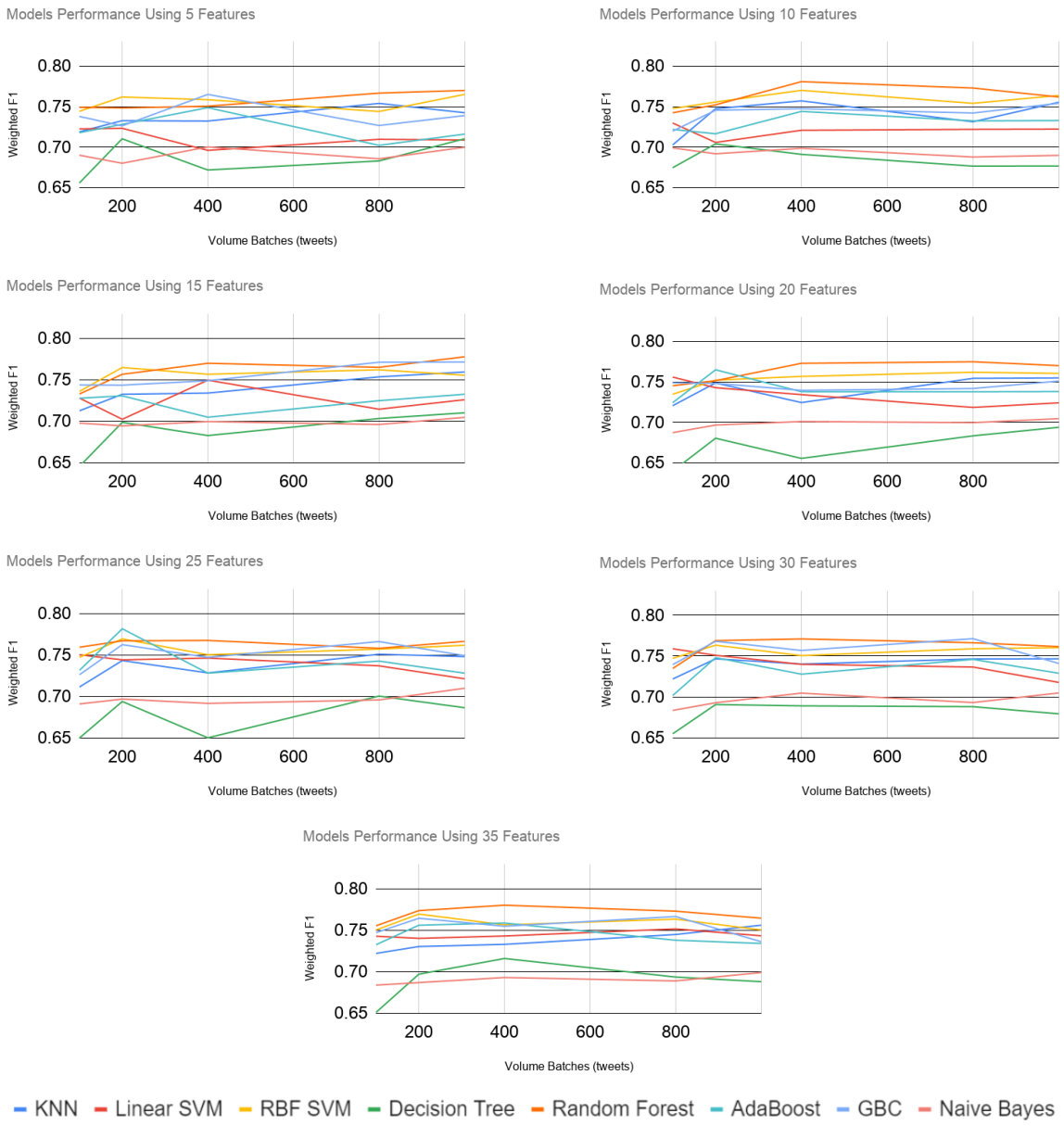


Figure 4.7: Performance of the different models with the variation of the volume batches in the selected features set. Maximum baseline performance is 0.53 which is not presented to allow better visualization of the differences in performance of the other models

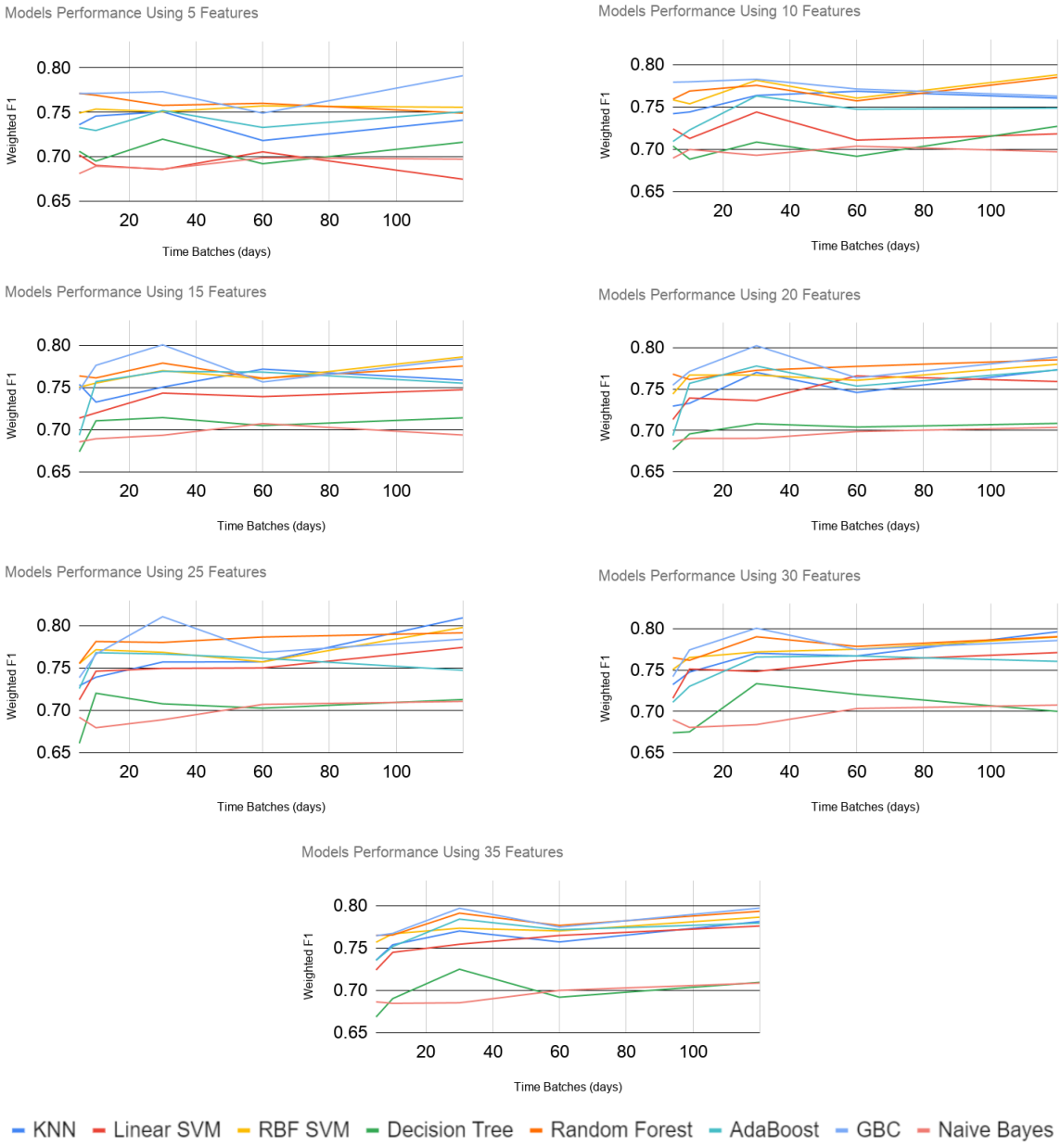


Figure 4.8: Performance of the different models with the variation of time batches in the selected features set. Maximum baseline performance is 0.53 which is not presented to allow better visualization of the differences in performance of the other models

Table 4.8: Cross-batch evaluation regarding each set of features for the highest performance model in each volume batch. The color scheme represents the weighted F1-score (highest in green) for the best model in each batch.

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.749	0.749	0.753	0.746	0.742
200	0.759	0.762	0.770	0.758	0.763
400	0.760	0.742	0.765	0.749	0.747
800	0.725	0.742	0.751	0.767	0.751
1000	0.721	0.740	0.747	0.751	0.770

(a) 5 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.748	0.757	0.751	0.754	0.759
200	0.752	0.756	0.756	0.754	0.754
400	0.750	0.749	0.781	0.757	0.752
800	0.731	0.761	0.770	0.773	0.763
1000	0.757	0.759	0.759	0.762	0.764

(b) 10 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.744	0.729	0.739	0.741	0.736
200	0.754	0.765	0.759	0.756	0.755
400	0.755	0.753	0.770	0.767	0.764
800	0.735	0.766	0.766	0.771	0.753
1000	0.754	0.766	0.778	0.771	0.778

(c) 15 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.756	0.731	0.747	0.752	0.754
200	0.742	0.765	0.770	0.750	0.739
400	0.758	0.755	0.773	0.767	0.769
800	0.750	0.756	0.774	0.775	0.766
1000	0.767	0.761	0.767	0.780	0.770

(d) 20 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.760	0.755	0.756	0.770	0.763
200	0.759	0.782	0.742	0.745	0.750
400	0.764	0.753	0.768	0.775	0.778
800	0.749	0.758	0.765	0.766	0.764
1000	0.769	0.761	0.769	0.759	0.767

(e) 25 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.759	0.728	0.751	0.752	0.755
200	0.788	0.769	0.760	0.773	0.765
400	0.776	0.765	0.771	0.775	0.770
800	0.743	0.743	0.761	0.771	0.757
1000	0.755	0.752	0.759	0.770	0.762

(f) 30 features

volume batch trained	volume batch tested				
	100	200	400	800	1000
100	0.756	0.729	0.758	0.762	0.765
200	0.764	0.774	0.763	0.766	0.760
400	0.754	0.758	0.780	0.775	0.771
800	0.747	0.756	0.765	0.773	0.773
1000	0.763	0.753	0.768	0.783	0.765

(g) 35 features

Cross-Batch Evaluation To understand the capabilities of the best models in handling with different batches, we performed a cross-batch evaluation for each set of batches and features within the time and volume experiments. Therefore, the best model for each batch and set of features, was used for a cross-batch evaluation.

The second part of this evaluation was conducted using a 5-fold cross-validation on the training dataset. The performance (measured using weighted F1-score) is presented in Table 4.8 for the volume-based batches and in Table 4.9 for the time-based batches.

The results related to the volume-dependent batches show that there are small variations in the performance of each model concerning the batch in which they were trained and the batches in which they were tested (the standard deviation ranges from 0 to 0.02 in all volume batches and features). Additionally, the models trained in one batch often do not achieve

Table 4.9: Cross-batch evaluation regarding each set of features for the highest performance model in each time interval batch. The color scheme represents the weighted F1-score (highest in green) for the best model in each batch.

time batch trained	time batch tested				
	5	10	30	60	120
5	0.771	0.768	0.755	0.761	0.778
10	0.740	0.771	0.761	0.767	0.760
30	0.715	0.750	0.773	0.765	0.773
60	0.741	0.733	0.750	0.760	0.765
120	0.719	0.731	0.752	0.769	0.791

(a) 5 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.779	0.748	0.771	0.785	0.783
10	0.759	0.780	0.747	0.762	0.772
30	0.760	0.778	0.783	0.790	0.798
60	0.719	0.748	0.764	0.771	0.771
120	0.760	0.768	0.778	0.782	0.788

(b) 10 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.764	0.761	0.764	0.784	0.785
10	0.760	0.776	0.763	0.784	0.773
30	0.745	0.766	0.801	0.798	0.796
60	0.754	0.757	0.780	0.772	0.779
120	0.768	0.776	0.777	0.778	0.786

(c) 15 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.768	0.766	0.774	0.793	0.777
10	0.761	0.771	0.760	0.775	0.781
30	0.748	0.780	0.802	0.788	0.798
60	0.769	0.757	0.769	0.777	0.768
120	0.763	0.781	0.778	0.776	0.789

(d) 20 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.755	0.768	0.764	0.777	0.778
10	0.770	0.781	0.775	0.776	0.792
30	0.745	0.778	0.811	0.799	0.804
60	0.772	0.757	0.765	0.787	0.775
120	0.770	0.768	0.774	0.788	0.809

(e) 25 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.765	0.768	0.784	0.782	0.781
10	0.778	0.774	0.779	0.768	0.777
30	0.763	0.784	0.801	0.786	0.788
60	0.774	0.774	0.791	0.779	0.790
120	0.773	0.786	0.788	0.790	0.796

(f) 30 features

time batch trained	time batch tested				
	5	10	30	60	120
5	0.765	0.773	0.778	0.782	0.790
10	0.768	0.767	0.766	0.778	0.771
30	0.756	0.764	0.797	0.776	0.781
60	0.765	0.764	0.770	0.777	0.790
120	0.764	0.770	0.782	0.779	0.797

(g) 35 features

the highest performance when tested in the same batch. Furthermore, independently of the batches in which the models were trained, the best results were located on average between batches 400 and 800. It is also possible to observe that models trained in the 100 tweets batch tend to show a slight increase in performance when evaluated in higher batches, especially in the 10-, 25-, and 35-feature sets. When analyzing the cross-validation results from the time interval batches, we can see that although the variation is small, the majority of the models seem to have an increase in performance when tested in batches with higher time intervals, regardless of which time-batch the model was trained in. This phenomenon becomes more evident as the number of features increases, with the majority of the best results for the 15, 20, 25, 30, and 35 feature datasets located on the right side of the respective tables.

4.2.7.3 Validation in Random Batches

The last step of the workflow is to simulate a real-world scenario by using never-before-seen data (validation dataset) to evaluate the selected approaches. We divided this evaluation into two steps. In the first step, we evaluated the performance of the best model that only uses account features. In the second step, we focused on the remaining models (time and volume batches). We chose this way of evaluation because: 1) it is infeasible to evaluate the models where the number of batches is different from zero in a batch=0 scenario since content features cannot be applied, and 2) since the validation set has a smaller number of entries, we prefer to present the evaluation separately instead of creating an ensemble and evaluating each batch in a smaller subset.

Account-based Models For this validation, we used the model that performed best in the batch=0 experiments (i.e., no tweet information was used to extract features). This model is Random Forests with a set of 30 features (although the set of 35 features yields similar results). The weighted F1-score obtained in cross-validation using the training set was 0.71.

We re-trained the model with the same parameters using the full training set and evaluated with the validation set. Using the same metric (weighted F1-score), the obtained results had slightly decreased to 0.69. However, considering that this evaluation is conducted in unseen data and without information from the account content, the selected model can still show a competitive performance compared to models using content-based features for the classification task.

Content-Based Models In this section, we evaluate the performance of the different models obtained in the individual and cross-batch evaluation using the validation dataset. We also present our solution to deal with the different number of tweets by applying 3 models that adapt to the volume and time interval of tweets in each account (ad_volume, ad_time, and ad_vt). These models were selected based on the results provided in Section 4.2.7.2 and 4.2.7.2 and were re-trained using the entire training set. The results are presented in Table 4.10.

As can be seen in Table 4.10, the individual models (ind_best_volume and ind_best_time) suffer from a drop in performance compared to the results obtained in Section 4.2.7.2. This is likely related to testing on unseen data, as well as the variation of the volume of tweets. In addition, the cross-batch models (ind_cross_volume and ind_cross_time) also show a decay in performance compared to the evaluation performed in Section 4.2.7.2 (although they show very similar performances in that evaluation). Finally, the results obtained by the adapted models (ad_volume, ad_time, and ad_vt) significantly outperform the results of the individual

Table 4.10: The results of the different models in the validation dataset using 5 different iterations of randomly generated batches for each entry considering 3 different types of distribution. The metric used is the average weighted F1-score for the 5 iterations with the standard deviation presented between brackets. The adapted models (ad_volume, ad_time, ad_vt) outperform the baseline models (ind_best_volume,ind_best_time) as well as the models that score higher on cross-batch testing (ind_cross_volume,ind_cross_time).

Tweets Batch	Models Validation using weighted F1-score						
Distribution	ind_best_volume	ind_best_time	ind_cross_volume	ind_cross_time	ad_volume	ad_time	ad_vt
25%[100,200] 75%[400,800,1000]	0.699 (0.008)	0.676 (0.004)	0.673 (0.012)	0.678 (0.012)	0.885 (0.003)	**0.907 (0.007)**	0.893 (0.006)
equal probability	0.698 (0.006)	0.676 (0.011)	0.669 (0.018)	0.670 (0.008)	0.888 (0.012)	**0.910 (0.01)**	0.891 (0.019)
75%[100,200] 25%[400,800,1000]	0.695 (0.01)	0.670 (0.018)	0.658 (0.012)	0.663 (0.006)	**0.902 (0.009)**	0.887 (0.013)	0.895 (0.006)

and cross-batch models. Although the results of the three adapted models are very similar, the volume-based model seems to have a slight advantage when the distribution of tweets in the interval [100,200] is heavier (0.902). In the other scenarios tested, the time-adapted model achieves slightly better results. Curiously, although the time- and volume-based adapted models have the lowest variation of results in the three scenarios tested, they do not achieve the highest performance in any case. One of the hypotheses for this phenomenon is that the model might be adapting to the point of causing a slight overfit regarding the training data. However, a more detailed investigation is needed, and we intend to address this in future work with more data.

Nonetheless, these results provide evidence that in a more pragmatic scenario where the number of tweets varies from account to account, ensemble models that adapt to a specific volume or time interval of posts, achieve higher performance detecting reliable and unreliable accounts compared to more traditional approaches. These results, in combination with the results from Section 4.2.7.3, can contribute to a more accurate classification based on the amount of information available for each account at a given time.

4.2.8 Conclusions

In the second part of this chapter, we studied and targeted the automatic classification of unreliable accounts on social networks. Similar to Section 4.1 and moving towards a more pragmatic approach, we diverged from the traditional bot account classification task to a broader interpretation of the problem and presented a scenario where the volume of posts in social media accounts changes. Our experiments have led to important conclusions that can provide relevant information on how to address this problem.

Word-embedding vectors extracted from the content (tweets) posted by the accounts are the strongest features in both time and volume batches. In these cases, only one non-

word embedding feature is presented in time-based batches. When no tweets are present, traditional bot detection features play an important role in the detection of unreliable and reliable accounts.

The results of our experiments with volume-based and time-based batches have shown that more content or a longer time period, does not equate to a higher performance of the model in detecting account reliability. In fact, the models tend to achieve their best performance between the 200- and 400-volume batches and in the 30-day time batch. The cross-batch evaluation also suggests that in some cases, models achieve higher performance when tested in batches different from the ones in which they were trained in. However, further investigation is needed to understand if these results are robust since the variation in performance is very small in the current experiments.

In our validation step, we started by evaluating the best model that relies solely on the information from the accounts. The weighted F1-score obtained was 0.71 and although this value is lower than that of some works in bot detection (e.g., the authors in [21] achieve an F1-measure of 0.79 and 0.91 using similar account-only data), our models have the distinction of working with unreliable and reliable accounts overall. We therefore argue that the trade-off between lower performance and greater generalization of the problem is appropriate, mainly because we are convinced that the absence of content is more likely in human-operated accounts than in bot accounts, due to the general goals of the latter (i.e. automatic content distribution).

However, the main result of this section shows that in a real-world scenario where the evaluated accounts vary in terms of number of posts, solutions that can adapt to this number (either through the time-period they were published or overall amount of tweets) present an improvement in performance in the order of 20% compared to more traditional approaches. These results highlight the importance of versatile ensemble models in the account reliability classification task and provide a solid foundation for future work in this area.

4.3 Chapter Summary

The experiments developed in Section 4.1 and 4.2 present useful contributions towards a real-world account detection system. They are a knowledge-based approach that provides a more detailed and robust result in assessing the reliability of an account (since it is based on behavior of accounts in the past), and a machine learning-based approach that can accurately classify accounts based on the amount of content available. Although both methods are limited in scope (more details in Chapter 6) and the results are not uniform in the type of output provided (numeric in the knowledge-based approach, categorical in the classification-based approach), we believe they are an important contribution to close the gap between

more experimental case studies and real-world detection systems.

In the next chapter, we shift our focus to the detection of unreliable content. Nevertheless, we maintain the goal of bringing more experimental case studies closer to a real-world scenario. In particular, we focus on how current approaches for detecting unreliable content in social networks perform over time by conducting a longitudinal evaluation on different combinations of features and models. This investigation is essential to ensure that the models deployed in a real-world application perform robustly over time, even with the potential rise of new topics in reliable and unreliable content.

Chapter 5

Performance of Unreliable Detection Models in Twitter Posts Over Time

In this chapter, we focus on the detection of unreliable content in social networks. This topic has been extensively covered in the literature. However, the studies presented usually focus on specific events such as the 2013 Boston Marathon or the 2016 United States Presidential elections. Detection of unreliable content has also been conducted using stance detection towards specific rumors or topics. However, there are fewer studies that address the development and evaluation of unreliable content detection models in a long-term perspective and the potential real-world problems and limitations that may affect the performance of these models. Therefore, rather than focusing on a short time interval or event, we evaluate the performance of unreliable content detection models over time. Model longevity is an important issue when dealing with real-world applications. An unreliable content detection system that loses performance over time can have a large impact on the information passed to users and consequently, negatively affect their perception of certain events. Accordingly, when aiming at developing a real-world application to detect unreliable posts on social networks, it is crucial to investigate how time can affect the performance of the models.

5.1 Problem Statement and Proposed Solution

Several factors can affect the performance of unreliable post detection models over time. For example, the topics in unreliable and reliable news can change, which in turn affects the information disseminated on social networks. Since the problem gained mainstream attention in the 2016 US presidential election, several events have served as a motto for

the spread of fake news and unreliable content on social networks (e.g., Brexit, 2020 US presidential election, and COVID-19). Analyzing the current approaches to the problem, it becomes unclear how long the trained systems will last before they are affected (or if they are affected at all) by time dependency of some topics and context change in unreliable and reliable content. More specifically, two main components may be affected by the time dependency of the content: the importance represented by the input features or indicators (i.e., their ability to distinguish between reliable and unreliable posts), and the most suitable model for the task, as some models may perform better over time than others.

In Chapter 2, we described different sets of features used in the literature and the performance obtained by different models, with some of them achieving very different results depending on the study conducted. This lack of consensus on the performance of the different sets of features/models may be due to the limited data in which the evaluation is conducted. In other words, a particular combination of features may perform better on a particular event or on data over a short period of time. However, there is an overall lack of understanding of the importance of these features over time. Therefore, our first research question is as follows:

RQ1: Which groups of features are most important and more stable (i.e., feature importance is preserved) when time-ordered data are added? To answer this research question, we select different feature groups and evaluate their importance in the collected data by analyzing which feature groups better describe our target variable (i.e., the unreliable/reliable label) using tweets from 15-, 30-, and 60-day periods. This way, we can analyze feature importance across time batches to better understand whether more data cause variation in feature importance. Although we could use the entire dataset for this analysis, we limit our knowledge to the train data to mimic the behavior of a real-world scenario where future data is unknown. Furthermore, we make this decision to avoid impacting our second analysis.

Our second analysis concerns the evaluation of machine learning models in identifying unreliable content in social networks over time. In other words, given a fixed time interval of training data, which "traditional" machine learning models and ensembles used in the literature exhibit better longevity and are they affected by topic changes? More specifically:

RQ2:Does the performance of unreliable content detection models decrease over time with the variation of the topics discussed? We hypothesize that certain features and models may perform better than others over time. For example, training a model with a Bag of Words and evaluate it over time could decrease fast in performance due to the dynamics of the topics discussed in reliable and unreliable content, as the vocabulary used in training may lose relevance over time. Alternatives that rely less on domain-specific

words may perform better. For example, features such as text statistics (e.g., number of words, percentage of capital letters) or cues to sentiment and emotion are more likely to be associated with domain and topic independence due to the majority of lexicon-based techniques used in these approaches.

In addition, state-of-the-art methods such as Word2Vec can provide useful features from the text. In particular, Google pre-trained Word2Vec seems to be suitable for this task since it was trained on a news corpus containing 100 billion words. Moreover, the experiments conducted earlier in Section 4.2 have shown that Word2Vec features are highly significant in a context where multiple topics are discussed. In addition, a customized Word2Vec can also be a suitable alternative to capture the vocabulary used in unreliable content thus providing a better approach to the problem.

To assess the two research questions, we have developed an experimental setup considering different groups of features and models from the literature. However, we will be evaluating their performance over time in a dataset extracted during an 18-month period. This will allow us to evaluate the robustness of the models due to the emergence of new topics (e.g., COVID-19) as well as the confidence in the results obtained due to the large time span of the dataset.

5.2 Experiment Workflow

To evaluate the longevity of the models over time, we developed an experimental setup whose different components are described in the following subsections.

5.2.1 Data Extraction

We begin by using the data from Twitter, which we extracted using the methodology described in Chapter 3. In this experiment, our data consists of tweets from July 22, 2019 to January 18, 2021. We also removed retweets from the sample and balanced the classes by day, i.e., on each day, we include the same number of reliable and unreliable tweets. The reasons for this decision are the following: Retweets are removed because of the impact they can have on the performance of the models. More specifically, retweets can occur in both training and test data. Therefore, retweets captured multiple times can have a positive impact on model performance (retweets on the test set are accurately classified due to their presence on the training set) or a negative one (models are trained with a large number of retweets and fail to capture the diversity of both classes). Concerning the balancing of the two classes, we chose this approach due to the imbalance artificially generated by the different number of sources in both classes. In addition, in a real-world detection system, depending on the user's

behavior and connections, the system may face different scenarios concerning the number of reliable and unreliable publications. Therefore, for this experiment, we decided to balance both classes. The final dataset consists of approximately 618k tweets with 309k tweets for each class.

Since our goal is to evaluate the longevity of the models, we use time intervals to distinguish our training data from our test data. Specifically, we assume a time series perspective in splitting the data where the training data and test data are chronologically ordered and separated by a time gap. In addition, rather than focusing on the percentage of data for each training and testing set, we split the data using days as our division unit. Therefore, we experimented with three different time intervals for the training data: 15, 30, and 60 days. Furthermore, we introduced a 15-day data gap between the training and test data to minimize the impact of very similar tweets in both datasets. A representation of the three scenarios considered is presented in Figure 5.1.

Figure 5.1: The three different scenarios considered concerning the size of the training set. Each square represent a chunk of 15-day data. Data are organized chronologically. Blue squares represent training data, grey squares represent data that are ignored (gap), and red squares represent testing data.

5.2.2 Feature Extraction

As mentioned earlier, we implemented several features from the revised literature to understand their time dependency and importance for detecting unreliable content. We chose to exclude social-based features due to the limitations of the extraction method (i.e., posts are extracted at an early stage and social interactions are still limited). We also chose to exclude

account indicators because of the large number of repeated accounts in our data over time and to avoid having the models distinguish between unreliable/reliable accounts instead of publications.

We divided the features into the following groups considering their main characteristics and the size of each group.

Bag of Words. The first set of input features uses a binary Bag of Words model to represent the text of each post. This set is intended to be used as the baseline features for the classification task.

Word2Vec (Google pre-trained model). Google Word2Vec model is a state-of-the-art approach that has been used in several studies and consists of about 1 billion words extracted from news articles. The large corpus in which it was trained and the large vocabulary it contains may be more accurate in capturing additional context than other approaches.

Word2Vec (Custom FakeNewsCorpus model). To supplement the lack of recent information on Google pre-trained Word2Vec and to include examples of unreliable content in the training data, we created a custom Word2Vec model using the FakeNewsCorpus dataset. Succinctly, the dataset consists of 1 million news articles from both fake and real sources (for more information see [222]). We chose to use only the headlines of each article to build the model, as including all articles from the dataset would be very computationally intensive. It is also important to highlight that the headlines used in the training of this Word2Vec model are dated prior to the tweets used in the 15-, 30-, and 60-day batches.

Lexical categories (Empath). Empath [73] is a tool similar to LIWC [226] that extracts lexical categories from texts. However, instead of simply relying on a lexical approach, it learns word-embeddings from 1.8 billion words. More specifically, Empath uses fictional stories to create word embeddings. Then, using a small seed of categories and a small number of root terms, it extends each category based on the similar terms provided by the word model. The main difference between Empath and other approaches is that, unlike Word2Vec and similar to LIWC, these categories are later validated by human annotators who remove the unfitted terms in each class. In addition, Empath presents 200 categories with highly correlated results (0.906 average Pearson correlation) with LIWC, making it a suitable and more accessible alternative.

Context-free features. This set refers to features extracted from the text but not necessarily to domain- and time-specific terms or words. We hypothesize that analyzing features

that are more independent of context and topics and not based on language models, may provide an advantage to the performance of the models over time. The features used in this group include text statistics (e.g., the number of words and exclamation points and the proportion of capital letters), parts of speech (e.g., the percentage of pronouns, nouns, and verbs) as well as sentiment and readability features.

5.2.3 Feature Analysis

To evaluate the impact of the different feature groups in the task of unreliable content detection on social media, we extracted features for each group in a 15-, 30-, and 60-day window. The following steps were then applied to each of the datasets to eliminate redundant features. First, features with a variance of less than 10% were removed. Second, features that had a correlation value superior to 90% were also discarded. To evaluate the importance of the features in the different scenarios, we applied the mutual information score. This score measures the dependency between the evaluated features and the target variable. The higher the score, the greater the dependency. Also, using this score, we included an additional group of features consisting of the best features of the different groups. Combining the data with the different groups of features results in 18 different datasets to be evaluated.

5.2.4 Models and Evaluation

Concerning the models and evaluation, we selected some of the most common examples in the literature, already mentioned in Chapter 2. Therefore, we used SVMs (radial and linear kernels), Decision Tree, Naive Bayes, and K-Nearest Neighbors (KNN). In addition, we complemented this selection with the following ensemble models: Gradient Boost Classifier (GBC), Random Forest, and AdaBoost. We also added a random baseline model (i.e., a model that randomly selects a class) for each of the experiments. As for the evaluation metric, we again rely on the weighted F1-score (see Equation 4.8).

5.3 Results

In this section, we present the results of the feature analysis and the different performances of each feature/model combination.

5.3.1 Feature Importance

Concerning feature importance, Figure 5.2 illustrates the 30 most important features in the different sets described in Section 5.2.2, for the 15-day training window. Analyzing the

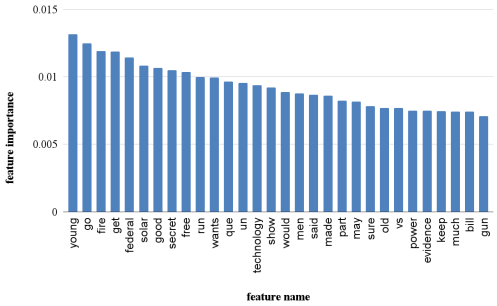
most important features in the lexical and context-free sets, we can see that the presence of some features can be related to earlier findings on the state of the art. More specifically, in the analysis of the main lexical categories extracted by Empath, some categories such as anticipation, emotional, and politics emerge. Similarly, context-free features also include the compound sentiment feature, which is an aggregation of sentiment scores [116]. The presence of emotion-related categories can be linked to how several studies have associated a high emotional (often negative) tone with unreliable information to influence users perceptions [211, 247, 112]. Consequently, due to the analysis and use of these types of features in several studies, it was likely that they would appear in this set.

In addition, other categories such as terrorism, power, dispute, and journalism in the Empath feature set, as well as some word features in the Bag of Words set (federal, power, evidence, bill, gun) may provide some cues on the importance of particular topics/terms to the task. Nevertheless, the mutual information score obtained by the best features in these sets ranges from 0.005 to 0.013, indicating that the correlation between these features and the target variable is weak. In fact, due to the analysis of the terms by unreliable and reliable accounts, as well as the importance of the features presented in the previous chapter, these results are not surprising. Since unreliable and reliable accounts tend to discuss the same overall topics and have low feature importance using Empath, it is plausible that the results are similar for the detection of reliable and unreliable posts.

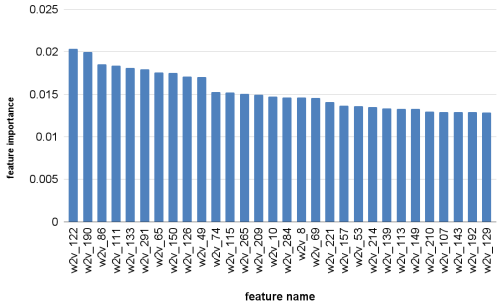
Better importance scores are seen in the word embeddings (Fake and Google) and context-free approaches (albeit with a very limited set), with the best features in these sets ranging from 0.02 to 0.03. Again, the results are similar to the work conducted in the last chapter, where Word2Vec features outperformed all other types when assessing the reliability of an account.

To better understand whether the low feature scores were due to insufficient data, we performed the same analysis in the second (30-day training window) and third (60-day training window) scenarios. The full charts for each scenario are presented in the Appendix A. In Table 5.1, we present the top 5 most important features for each feature set in each scenario for comparison purposes.

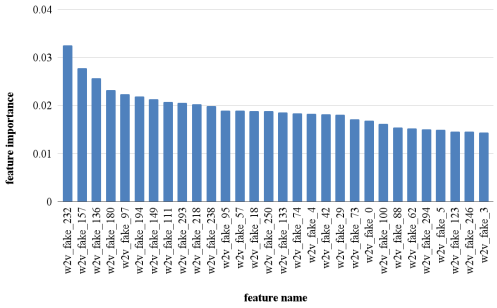
(a) Bag of Words



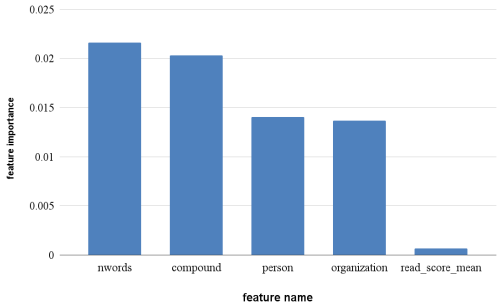
(b) Google Word2Vec



(c) Fake Word2Vec



(d) Context-free features



(e) Lexical Categories (Empath)

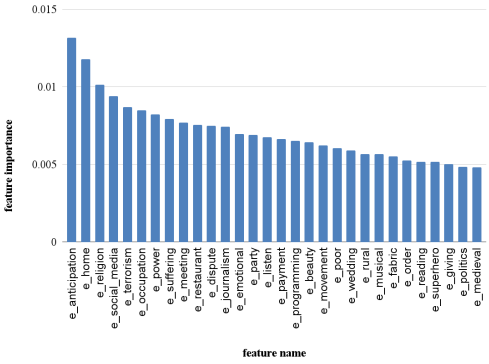


Figure 5.2: Importance of 30 most relevant features from different feature sets using a 15-day interval. In some scenarios, due to preprocessing applied (removal of low variance and highly correlated features), total number of features is inferior to 30.

Table 5.1 shows how the top features from each set change with more data. In particular, we can see that bag of words features are completely distinct in the three scenarios considered. The results are similar in the lexical categories feature set.

The Word2Vec sets show some similar top features in the 3 scenarios considered. The Google pre-trained model presents similar features between the 30 and 60 days scenarios, namely vectors 43 and 61. The Word2Vec trained with FakeNewsCorpus titles not only shows similar

vectors in the different scenarios but the top feature is the same in the 15-, 30-, and 60-days scenarios. In addition, the results provided by these feature sets are higher than the lexical and Bag of Words sets, which may lead to a better performance of the models trained with these input features.

Similar to the word embeddings, the context-free features also present equivalent scores. However, the removal of low variance and highly correlated features leads to a very small and similar set across all 3 scenarios considered. Moreover, although the number of words is the most important feature in the first scenario, the compound sentiment and the number of entities are ranked higher in the third scenario. However, this is not due to a higher overall ranking of "compound" and "person" features (i.e., the number of people identified in the text), but to the loss of importance of the "nwords" feature (number of words in the text). In addition, sentiment indicators are also present in this set of features, showing the importance of this type of feature in the task of unreliable information detection.

In summary, some important conclusions can be drawn from the analysis of the most prominent features for each set. First, similar to the experiments conducted in Section 4.2, the sets with word embedding features achieve some of the highest feature scores. In addition, the context free set, composed of more "traditional" features and more independent of the context of the posts, also present similar importance scores. Second, although these sets achieve the best feature scores, they are still not ideal and show a low correlation between them and the target variable. Nevertheless, each feature is evaluated independently, and therefore its importance may vary when they are combined in the training of a model. Third, when considering the best feature sets (word embeddings and context-free), we can see that the top features are stable in the three scenarios considered, since they retain some features between sets and the score do not deteriorate with the introduction of additional data. Similar to [112], we hypothesize that this may lead to robust performance of the models over time and that performance degradation occurs slowly. In the study by Horne et al. [112], performance degradation was observed after 38 weeks in a similar scenario, but it was related to articles rather than posts on social media.

Therefore, to answer our RQ1, considering all sets of features, we can conclude that feature importance is not constant as more data are added. However, the word embeddings achieve higher feature importance than the other feature groups and maintain some of the top features over the 15-, 30-, and 60-days time batches. Therefore, we hypothesize that these feature groups are more likely to remain unaffected by the change of topics in unreliable and reliable content and consequently will guarantee better overall performance of the models.

Table 5.1: Feature importance scores regarding different set of features and three scenarios considered (15-, 30-, and 60-day windows).

15 days		**30 days**		**60 days**	
Bag of Words					
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**
young	0.01316	away	0.01001	please	0.00668
go	0.01249	police	0.00967	look	0.00616
fire	0.01190	found	0.00830	freedom	0.00493
get	0.01189	violence	0.00781	ilhan	0.00481
federal	0.01142	political	0.00752	face	0.00467
Word2Vec (Google pre-trained)					
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**
w2v_122	0.02034	w2v_43	0.02159	w2v_43	0.02609
w2v_190	0.01997	w2v_61	0.01947	w2v_192	0.02051
w2v_86	0.01854	w2v_119	0.01802	w2v_113	0.01988
w2v_111	0.01838	w2v_271	0.01733	w2v_61	0.01868
w2v_133	0.01812	w2v_149	0.01729	w2v_119	0.01843
Fake Word2Vec					
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**
w2v_fake_232	0.03256	w2v_fake_232	0.03280	w2v_fake_232	0.03402
w2v_fake_157	0.02777	w2v_fake_73	0.03087	w2v_fake_73	0.03396
w2v_fake_136	0.02564	w2v_fake_29	0.02923	w2v_fake_29	0.02990
w2v_fake_180	0.02317	w2v_fake_250	0.02635	w2v_fake_136	0.02736
w2v_fake_97	0.02238	w2v_fake_157	0.02599	w2v_fake_1	0.02558
Lexical Categories					
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**
empath_text_anticipation	0.01317	empath_text_stealing	0.00863	empath_text_internet	0.00995
empath_text_home	0.01178	empath_text_driving	0.00660	empath_text_messaging	0.00742
empath_text_religion	0.01016	empath_text_hipster	0.00650	empath_text_disgust	0.00676
empath_text_social_media	0.00940	empath_text_fun	0.00649	empath_text_journalism	0.00571
empath_text_terrorism	0.00871	empath_text_timidity	0.00629	empath_text_economics	0.00506
Context Free Features					
Feature	**Score**	**Feature**	**Score**	**Feature**	**Score**
nwords	0.02162	compound	0.02510	compound	0.02134
compound	0.02034	nwords	0.02137	person	0.01658
person	0.01404	person	0.01478	nwords	0.01648
organization	0.01369	organization	0.00771	organization	0.01186
read_score_mean	0.00066	read_score_mean	0.00048	read_score_mean	0.00208

5.3.2 Models' Evaluation

In this section, we evaluate the performance of the different models using the feature sets described previously. For models' training, we consider the three different scenarios. For the evaluation of the models, we use the remaining data (except for the 15-day gap).

Similar to what was done in the previous section, we chose to remove low variance and highly correlated features. However, we did not use the mutual information score to determine a fixed number of features. Instead, we added a new group composed of the best 15 features (according to the mutual information score) in all groups considered.

The results for the first scenario are presented in Figure 5.3.

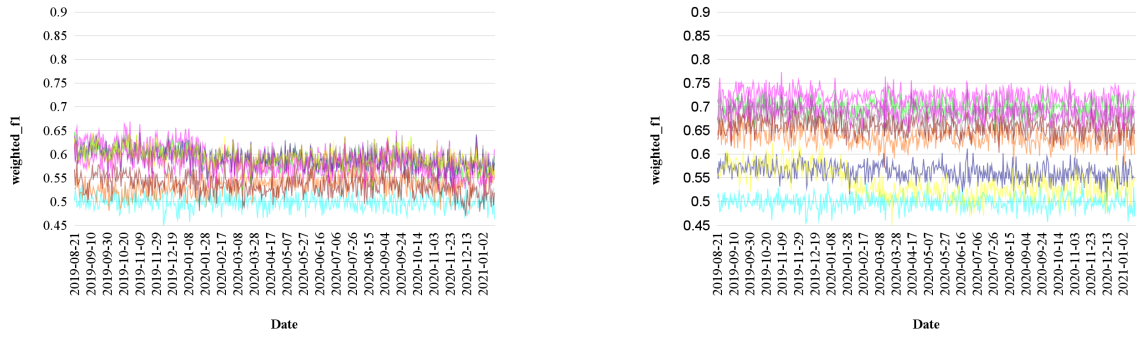
In the first scenario analyzed, it is clear that some sets of features perform better than others (using a training window of 15 days). The Bag of Words, context-free and lexical categories features are the ones that achieve the lowest values over time. The performance of the best models ranges between a weighted F1-score of 0.6 and 0.65 for the context-free features and 0.55 and 0.65 for the Bag of Words and lexical categories. The Google and Fake Word2Vec sets perform better with the best models averaging between 0.65 and 0.75. Finally, the models trained with the best 15 features of all sets achieve similar performance. However, the best model in this set achieves slightly better performance, with the lower bound of being around 0.7.

As for the best models, Support Vector Machines (RBF and Linear) achieve the best results on average, except for the set with the best features, where Random Forests and GBC (0.71 and 0.72, respectively) slightly outperform Linear SVM (0.70).

The feature set used has a large impact on the longevity of the models. When using the Bag of Words feature set, the best model (on average) is the RBF SVM. However, we can see that after approximately 5 months, the model's performance begins to decline. The same can be observed for the context-free features, where most models show a steep drop in performance, with the Linear SVM being the only model that maintains its performance over time. As it was hypothesized, feature sets that use word embeddings appear to better maintain the performance through the test set. Finally, the models with the highest performance on the best feature set show similar results to the feature sets with word embeddings. This is likely due to the features common to both approaches, as most of the best features are reused from the word embedding sets. Nevertheless, the best model (GBC) shows a slight drop in performance over time, with RBF-SVM, KNN, and Naive Bayes showing a sharp drop in performance between January and February 2020. The performance drop in this interval is also observed for the RBF-SVM model in the Bag of Words feature set and for the majority of the models using context-free features.

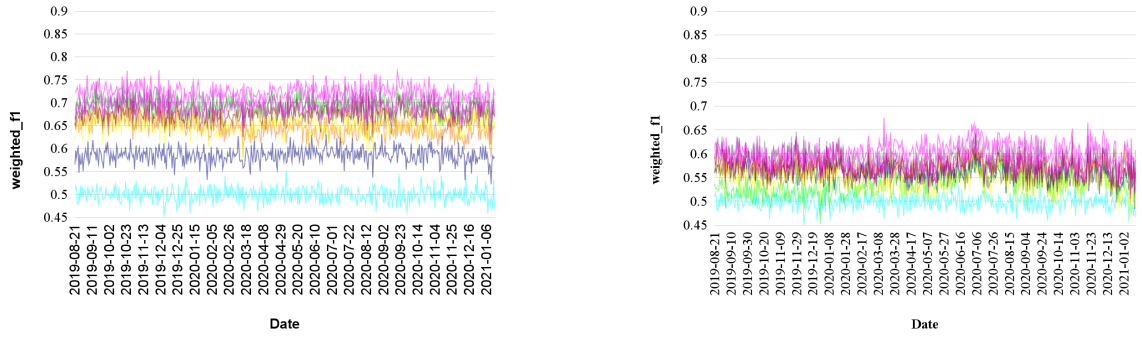
(a) Bag of Words

(b) Google Word2Vec



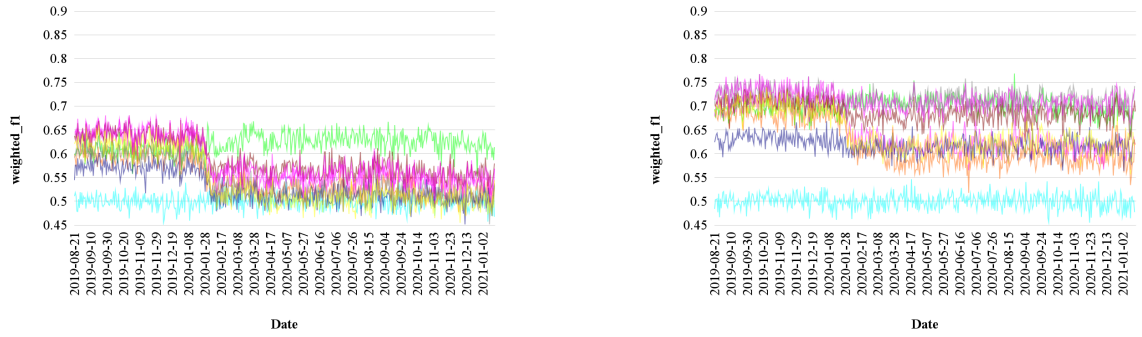
(c) Fake Word2Vec

(d) Lexical Categories (Empath)



(e) Context-free features

(f) Best 15 features



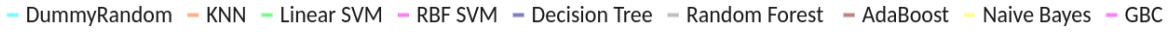


Figure 5.3: Performance evaluation (using weighted F1-score) of different models for each feature set.

Our hypothesis to explain the sudden drop in performance in some models is thus related to the drift of topics being discussed in both reliable and unreliable content. In fact, the report on the first COVID cases in the United States [40] and Europe [7] matches the time interval during which some models lose performance. Thus, we hypothesize that in some cases concept drift (more precisely, virtual concept drift) may affect the performance of the models over time. Nevertheless, models that use word embedding vectors as input features

seem to be more stable and suffer less from the sudden change of topic.

We proceed to analyze the second and third scenarios and investigate how the models trained with different feature sets are affected by more extensive training windows (and consequently, larger volumes of data). The results with the different features/models in these scenarios are presented in the Appendix B. To summarize the results achieved, in Figures 5.4 and 5.5 we focus on word embedding features sets since they achieve the best overall performance. Furthermore, for comparison purposes, we only present the best model for each feature set.

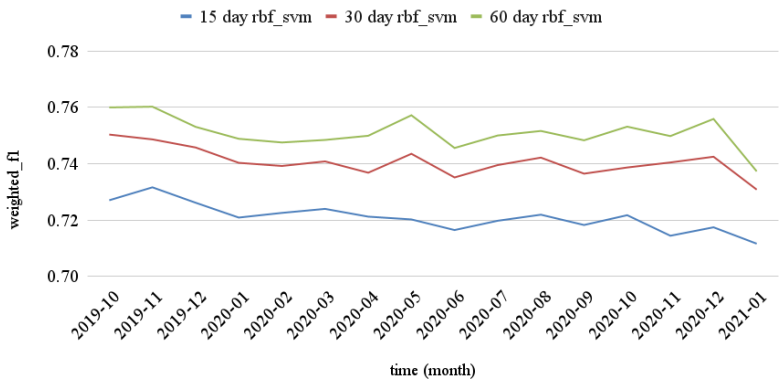


Figure 5.4: Comparison of best models for each scenario (initial training window of 15, 30, and 60 days), averaged by month, using features extracted from Google pre-trained Word2Vec.

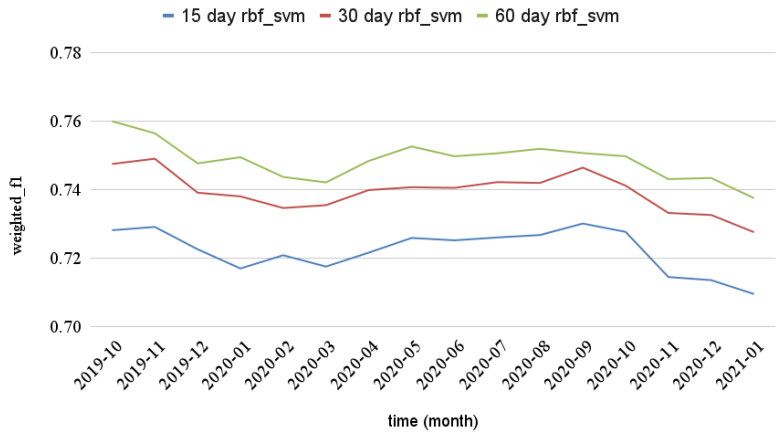


Figure 5.5: Comparison of best models for each scenario (initial training window of 15, 30, and 60 days), averaged by month, using features extracted from fake Word2Vec.

In both cases, we can observe that the models trained with 60-days data perform slightly better than the others. In fact, as the volume of data increases, so does the performance of the model. The best average models in both feature sets are SVM-RBF. Therefore, adding

data from subsequent days increases the overall performance of the best models.

In conclusion, to answer RQ2, some combinations of models and features have a performance drop over time (more specifically, the groups of context-free features and Bag of Words). However, the combination of word embedding features and SVMs seem to perform the best. In addition, these models do not seem to be affected by the change of topics in reliable and unreliable news, specifically with the rise of the discussion around COVID-19. Similar results were obtained in the case of news articles [112], where the performance decreased very slowly and changes in news concepts did not seem to affect the performance drastically.

5.4 Performance Boosting with BERT and RoBERTa Derived Features

According to our previous experiments, word-embedding sets performed the best among the different feature groups tested, as the best models trained with these features tend to maintain their performance over time and are less affected by topic changes. Therefore, it is important to further investigate and evaluate the impact of newer vector representations. To this end, we select the BERT and RoBERTa models to extract new text representations and evaluate their impact on the task of detecting unreliable content. BERT and RoBERTa are both state-of-the-art pre-trained language models that further enhance the performance in various NLP tasks. There are two different ways to use these models. The first is to use BERT/RoBERTa as a base model and add additional layers to fine-tune for specific tasks. The second is to use these models as feature extractors since BERT and RoBERTa can output word embeddings that are useful features for different models. We chose the latter to better understand how these feature sets affect the traditional machine learning model used in the previous experiments. Thus, we added four sets of features based on the performance of the word embeddings in the task of detecting unreliable content. These sets are derived from the large and base models used in BERT and RoBERTa. The main differences between these models are related to the number of encoders and consequently the number of attention heads, parameters, and size of the output. As for the differences between BERT and RoBERTa, there are 3 main aspects. First, RoBERTa removes the next sentence prediction task from BERT, uses more data and larger batches to train the model, and applies dynamic masking patterns to the training data [140]. The main differences between the 4 pre-trained models are listed in Table 5.2.

We apply the following configurations to the 4 models to extract embeddings from tweets. First, we define the maximum sequence length for the models based on the analysis of the number of tokens in a sample of the training data. Figure 5.6 shows the distribution of tokens in tweets. Other configurations include the batch size (set to 32 in this experiment) and the

Table 5.2: Main differences between the 4 BERT-derived models used for feature extraction

	BERT-base	**BERT-large**	**RoBERTa-base**	**RoBERTa-large**
Parameters	110 million	336 million	110 million	336 million
Layers of Transformers	12	24	12	24
Attention Heads	12	16	12	16
Hidden States	768	1024	768	1024
Data	BookCorpus + English Wikipedia (3 300 M words)		BERT data + CC-News +OpenWebText+Stories (33 000 M words)	
Key Features	Bidirectional Transformer with Mask Language Model and Next Sentence Prediction		No Next Sentence Prediction Larger Batches Dynamic Masking	

internal validation split (set to 15% of the training set and considering the chronological order of the data). The results of using the output vectors of the different models derived from BERT as features for detecting unreliable and reliable content using a training window of 15 days are shown in Figure 5.7.

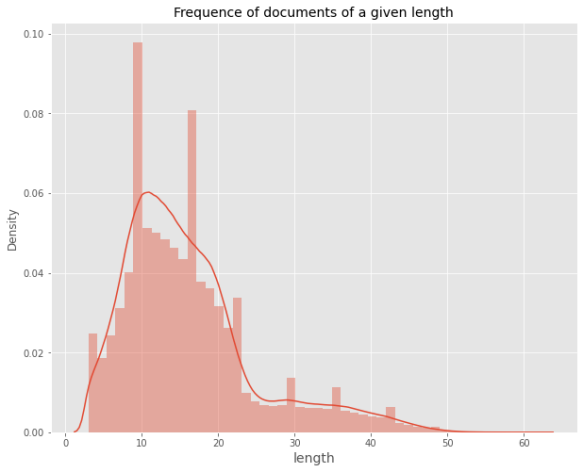
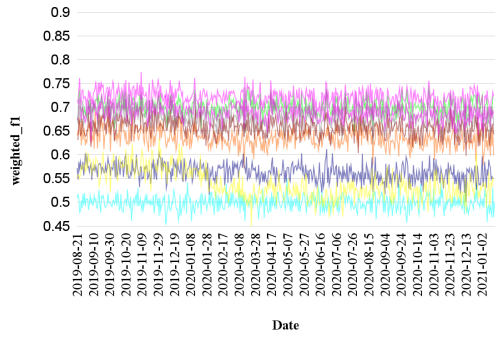


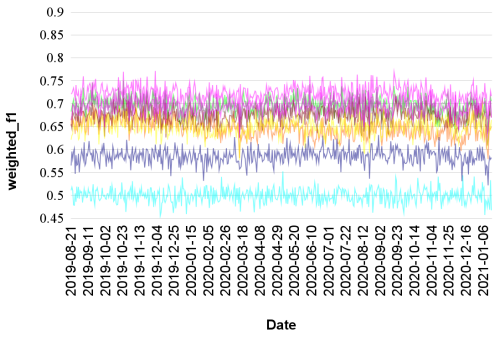
Figure 5.6: Distribution of tokens in tweets

Compared to the best feature sets from the previous experiment (Google Word2Vec and Fake Word2Vec), it is clear that the best models trained with BERT and RoBERTa derived features perform better than the best models trained with Word2Vec features. This is true not only for the 15-day scenario shown in Figure 5.7 but also for the 30- and 60-day scenarios (see Annex B.1). In addition, it is interesting to note that while the SVM still outperforms the other models, it is the Linear kernel that performs the best in contrast to the radial basis kernel from the SVM trained with Google and Fake Word2Vec embeddings. Another interesting result related to the features derived from BERT is that the best BERT-base model obtains a higher average F1-score than BERT-large (0.795 and 0.745, respectively), contrary to other NLP tasks where the larger version of the model tends to outperform the base model. Although most studies indicate an increase in performance when moving

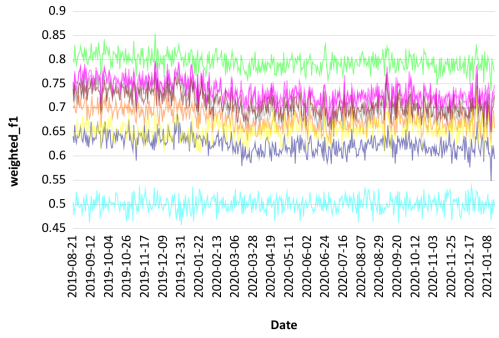
(a) Google Word2Vec



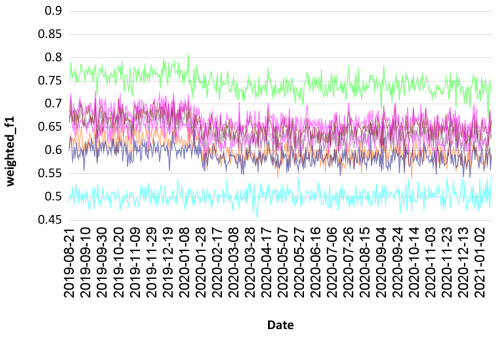
(b) Fake Word2Vec



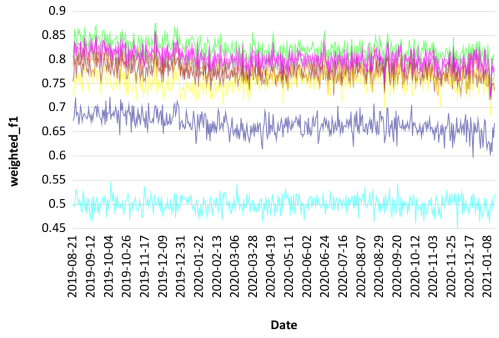
(c) BERT-base



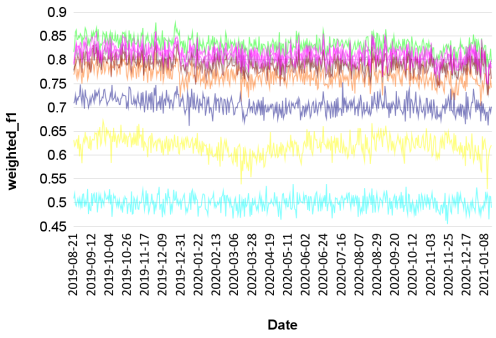
(d) BERT-large



(e) RoBERTa-base



(f) RoBERTa-large



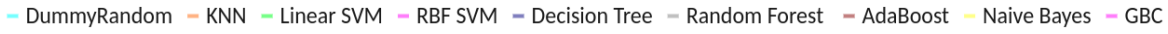


Figure 5.7: Performance evaluation (using weighted F1-score) of the different models for each BERT-derived feature set in the 15-day training scenario

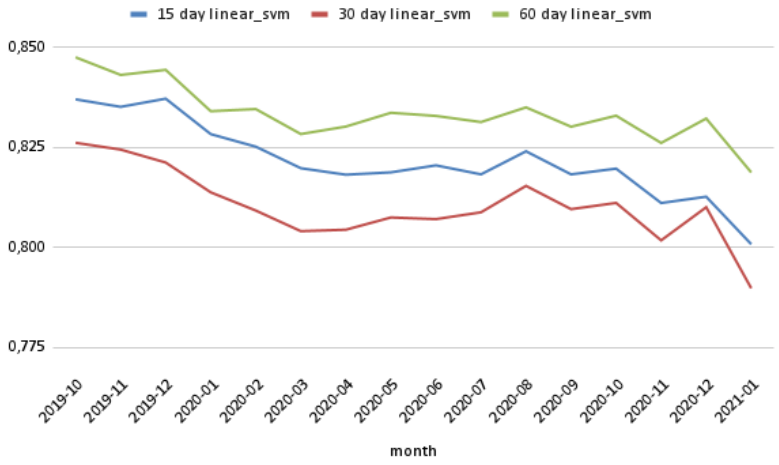


Figure 5.8: Comparison of the best models for each scenario (initial training window of 15, 30, and 60 days), averaged by month, using the features extracted from RoBERTa-base.

from BERT-base to BERT-large, the results in our experiment show a drop of about 0.05. Nevertheless, there is some evidence suggesting that, at instance-level, the performance of larger models may be affected by fine-tuning variance [265]. However, further investigation is needed to better understand this phenomenon in the context of this particular task and data.

Focusing on the last two feature sets, RoBERTa-base and RoBERTa-large, we can see that they further improve the performance of the best model with an average weighted F1-score increase of 0.03 and 0.04 (respectively) compared to BERT-base. Moreover, the performance difference between the best model and the others has decreased, with the average weighted F1-score being higher in the models with the features extracted from RoBERTa-base and RoBERTa-large. In addition, this increase remains stable over time and does not occur due to peaks on specific days. This proves once again the robustness of the models trained with BERT and RoBERTa-derived features.

Similarly to the previous section, we examine the effects of extending the training data from 15- to 30- and 60-days. Accordingly, we present in Figure 5.8 and 5.9 the performance of the best models in each scenario.

When we compare the best models for each scenario using RoBERTa-base and RoBERTa-large with those using Word2Vec features (Figure 5.4 and 5.5), we can see small differences. Apart from the performance improvement, additional data does not lead to an increase in the weighted F1-scores (unlike what happens in the Word2Vec feature-based models). More specifically, the best RoBERTa-base model trained with 30 days of data performs worse than the model trained with 15 days of data. In RoBERTa-large, the same scenario does not occur, except for a performance drop in December 2019 that temporarily puts the 30-

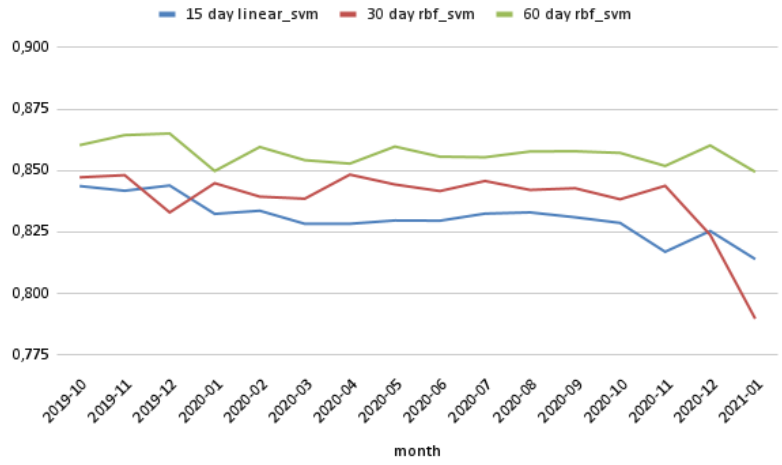


Figure 5.9: Comparison of the best models for each scenario (initial training window of 15, 30, and 60 days), averaged by month, using the features extracted from the RoBERTa-large.

day model performance below that of the 15-day model. Nevertheless, in both RoBERTa-base and RoBERTa-large, the model trained with the 60-day batch outperforms the best models trained with the 15- and 30-day batches. However, it is important to emphasize that the improvement obtained by the 15- and 60-day batch model is not significant (0.01 in RoBERTa-base and 0.03 in RoBERTa-large). Therefore, in a more pragmatic approach where human annotation is required and labeling and computing costs exist, achieving marginally better performance with significantly more data might not be worth.

5.5 Conclusions

In the previous experiments, we compared the performance of different sets of features/models in the task of unreliable content detection in social media. These were conducted in a setting where the evaluation was done over a longer and ordered period of time compared to related work from the current literature. We observe how some groups of features such as the context-free and Bag of Words sets are affected by topic drifting while models trained with word embeddings are more resilient to these changes. To the best of our knowledge, this is the first work to address the longevity of unreliable detection models in social networks and presents the evaluation using a dataset of tweets over a period of at least 16 months, making this the main contribution of this study.

The results obtained are mixed with respect to the expected outcome. Although we anticipated that the models using Word2Vec feature sets would perform better, it was unexpected that the models trained with context-free features had such a large performance loss, since they are more independent of the topics discussed (unlike Bag of Words, for example). These

results may be due to the small number of features in this set, which resulted from the removal of highly correlated as well as low variance features. In the current literature, the majority of experimental settings that use these features rely on cross-validation without chronological order (for example [39, 104]). Consequently, in this settings, important features can be assessed from the most recent tweets in the dataset (something that is not possible when considering chronological order in the training and test data). Hence, feature importance and selection are often considered time-independent in the literature. However, in a more pragmatic scenario, important information for these processes is not reachable. In addition, as it was also observed for fake news articles [112], the performance of the best models degrades very slowly (as can be seen in Figures 5.4 and 5.5).

Based on the good results obtained using word-embeddings, we also consider 4 additional feature sets based on BERT and RoBERTa pre-trained models encodings. Although these models have been previously used in stance detection tasks and certain cases of unreliable detection in social media (especially image manipulation), their use is overlooked in a problem like the one presented in this chapter. The use of these models as feature extractors resulted in robust and stable performance improvements, highlighting their importance for NLP-related tasks and, in particular, for the unreliable content detection task.

To summarize, the main contribution of this chapter is the evaluation of different combinations of models/features in an experimental setting more similar to a real-world scenario where models are trained in a specific time interval and "deployed" in a experimental environment where they are evaluated using posterior reliable and unreliable posts. In this setting, the combination of RoBERTa embeddings and SVMs outperforms all other approaches and thus seems most suitable to be considered in a real-world unreliable content detection system.

Chapter 6

Conclusions

The work presented in this thesis takes some of the groundwork of the current literature and adapts it to a more pragmatic scenario by imposing real-world constraints to two of the most important tasks related to unreliable information in social networks: the detection of unreliable content and the accounts that disseminate it.

Throughout this work, it became clear that introducing constraints to achieve a more realistic scenario degrades the overall performance of some proposals presented in the literature, mainly because they fail to generalize in adverse situations that can easily occur in the context of social networks and unreliable information. In particular, we can draw the following conclusions from the work carried out in this thesis.

First, assuming a scenario where unreliable and reliable accounts can be operated by both humans and bots, we develop metrics that can classify these based on prior knowledge. Analyzing the highest scoring accounts, we provide evidence that the metrics are able to correctly assess the behavior and impact of unreliable and reliable accounts. Furthermore, we show how the highest scoring unreliable accounts using our metrics are largely different from the highest scoring accounts identified by bot detection systems, suggesting the complementary nature of both approaches.

Second, the automatic classification of reliable and unreliable accounts according to the previous scenario increases the difficulty of the task (compared to bot detection), as common features such as posting frequency did not present similar importance. Moreover, the introduction of constraints on the content of each account also highlighted some shortcomings of current experimental approaches, where detection models trained in accounts containing a certain number of posts cannot be generalized to a more dynamic scenario where this number may vary.

Third, regarding the detection of unreliable information, longitudinal evaluation of different

combinations of features and models in the current literature has shown that not all proposals perform robustly over time, especially when there is a sudden change in the topics discussed.

Regarding the main contributions of this thesis in the two problems studied, we highlight the following:

- An extraction methodology was developed, complemented by a distant-labeling approach, to annotate large volumes of tweets over time. Moreover, unreliable accounts, as well as reliable and unreliable publications were analyzed in a preliminary exploratory approach.
- Knowledge-based metrics were proposed to characterize reliable and unreliable accounts based on their impact and behavior in social networks.
- The development of models to detect unreliable accounts that adapt to the content of each account. We show that these perform better than traditional approaches in a scenario where the volume of content in social media accounts can change.
- The longitudinal evaluation of different combinations of features/models taken from the state of the art in the task of detecting unreliable content. In addition, we introduce the use of BERT-based models for feature extraction in this particular context and show their superior performance and robustness over the analyzed time period.

Although we believe that the experiments conducted in these tasks make an important contribution to the development of a real-world application, it is important to highlight their limitations and strengths, as well as possible future directions to not only mitigate the current limitations of this work, but also improve upon the results presented.

6.1 Strengths and Limitations

In considering the whole process of data extraction and annotation, it is important to emphasize that our definition of unreliable information is broader than the definitions used in most works in the related literature. This allows us to extend the work in two directions. First, in evaluating whether an account is unreliable or not, we can also consider accounts that disseminate different types of unreliable content. Second, by using multiple types of unreliable social media publications, we can generalize our content detection models and identify different types of unreliable content. Another advantage of the extraction and annotation method discussed in Chapter 3 is the retrieval of large amounts of data without human intervention. This feature is important in the experiments conducted in Chapter 5 because it allowed evaluation over a longer period of time and in a larger set of publications

than the majority of works in the current literature, which improves the robustness and confidence of the results presented.

However, some limitations also apply to the distant annotation approach. In particular, it does not ensure that all labels are correctly assigned, as this process assumes that a tweet containing an unreliable/reliable link supports the type of content associated with that link. In reality, tweets may contain a link to an unreliable news article to alert other users to the disinformation it contains. On the other hand, posts from malicious accounts may also contain links to reliable news websites in order to criticize these news sources and accuse them of spreading false content. Furthermore, the website to source annotation also has some limitations, since it is not guarantee that all articles from a website labeled as unreliable will also contain unreliable information. In an ideal scenario, each extracted tweet should be validated by a group of human annotators/experts. However, due to the large number of tweets used in this work, the cost associated with this task would be impractical. Therefore, based on the success of related work (see [13, 174]), we opted for this distant-labeling approach, even though we recognize the limitations associated with it.

Another limitation in data extraction and annotation that is important for implementing a real-world application is the inclusion of a neutral class, as posts that disseminate personal or irrelevant information are not considered in current approaches. However, without a proper human-annotation process, it is difficult to label neutral posts, as they contain a variety of content. Users on social networks may discuss the food they eat, the music they listen to, games they play, family photos, travelogues, restaurant reviews, and so on. Although other approaches could be considered, ideally a large and diverse set of neutral posts would better integrate with the goal of this work.

In Chapter 4, we explore two solutions to the problem of assessing whether an account is reliable or not: a knowledge-based approach and a supervised-based approach. Both solutions can be combined into an unreliable detection system that can be used in a real-world scenario where the lack of knowledge about a particular account can be solved with a prediction-based strategy. One of the main factors that differentiate these solutions from the current state of the art is the study of the problem from the perspective of unreliable vs. reliable accounts instead of human vs. bot accounts. As mentioned earlier, we believe this is the more accurate way to tackle the problem, as there is growing evidence of the impact of human-operated accounts on the distribution of unreliable content.

In the first part of Chapter 4, we introduced metrics that can measure the impact of an account in a social network environment. In addition, we gained some important insights using the highest scoring accounts and a state-of-the-art bot detection system. First, the proposed framework and metrics are able to detect and classify unreliable accounts that may pose a threat to the Twitter ecosystem. Second, the presented metrics are useful for evaluating and scoring unreliable accounts that are not captured by bot detection systems

(such as Botometer), proving once again the need for an alternative to the bot vs. human approach. When examining the prevalence of unreliable accounts in Twitter, the results also show that unreliable accounts are responsible for 4% and 6% of tweets extracted through the Search and Stream APIs, respectively. However, when journalistically relevant and controversial keywords are used, this percentage increases to 25% and 36%, respectively. When evaluating the nature of unreliable accounts using our main metrics (IMP and IMP_{sf}), we concluded that searching (using the Search API) for controversial and relevant topics results in tweets from accounts that score higher in the proposed metric, which is problematic for Twitter users and the factuality of the content on this social network.

However, as mentioned earlier, there are some limitations to the extraction and annotation methodology used in this thesis. In particular, two limitations affect the proposed metrics. First, Twitter API rate limits force us to restrain the number of tweets extracted per day. This can have an impact on the number of unreliable posts captured and thus affect the score for each account. An ideal scenario would be to capture all daily tweets that contain an unreliable/reliable link. However, this would require access to a more restrictive API and could potentially escalate the resources and computation costs. The second limitation (which is also related to Twitter's API rate limits) is the lack of an update mechanism for each account score. Due to the dynamic nature of social feedback indicators, the posts may still have a low number of favorites/retweets at time of capture, and thus may not represent the actual impact they may have on the network. In addition, updating these posts may take some time due to the limited calls to the Twitter API.

In the second part of Chapter 4, we explored and targeted the classification of unreliable accounts in social networks. The main strength of the proposed approach is the evaluation of accounts in a dynamic setting where the content available may vary, thus showing a similar behavior to the real world. Moreover, the development of a model that adapts to the volume or time of posts proves to be more effective than current approaches in this scenario. However, there are also some limitations to the experimental results presented.

The first is the small dataset size which may limit the robustness and reliability of the results presented. Second, the limitation on the post-extraction process forces the potential score prediction (i.e., regression task) to be transformed into a classification task. Ideally, each post in an account timeline would be human-annotated. However, as previously mentioned, that would not be feasible within the scope of this work. Consequently, a more fine-grained and uniform numerical output is not possible in a prototypical application using knowledge and prediction-based approaches in the current scenario.

In Chapter 5, we focus on a more practical approach to the problem of identifying unreliable content in social networks by evaluating the performance of different sets of features and models on a dataset of 18 months of tweets. By splitting the training and testing data in chronological order, we were able to examine how the importance of features changes over

time and how each combination of features/models is affected specifically in events such as the Covid-19 pandemic.

By training content detection models on 15-, 30-, and 60-days batches of tweets and evaluating their performance over a time interval of more than a year, we can conclude that the performance of the models is the highest and most stable over time when word-embedding features are used. These results contribute to the development of a more pragmatic unreliable detection system that can be used in a real-world scenario. For example, a browser add-on capable of evaluating posts on social networks, where a user can enter a tweet and the application outputs a class based on the reliability of the content.

With respect to Chapter 5, we highlight some limitations due to the limited scope of the language and sources used in the experiments conducted. In particular, we focus on the English language and the sources retrieved from MBFC and OpenSources. It is difficult to assume that the same results obtained in this experiment (as well as in the experiments in Section 4.2) are generalizable to other languages/idioms. This is due to the performance of the required word-embedding models as well as the different syntax and grammatical rules. As for sources, we assume that the introduction of new sources would not severely affect the performance of the best models and that these would be able to generalize as well as they did with the introduction of new topics. This is mainly because we focus on the content in the tweets that these sources disseminate, rather than the articles in the sources themselves. In addition, Horne et al. [112] conducted experiments that measured the impact of new sources in models that detect reliable and unreliable articles and concluded that they are robust to this change. Given these two factors, we are led to believe that the models would generalize well with new reliable and unreliable sources.

Solutions to mitigate some of the limitations discussed, as well as some of the possible research lines are presented in the next section.

6.2 Future Work

In this section, we discuss some possible research steps to complement and improve the limitations present in this work.

In the data extraction and annotation methodology, two main steps can be taken to improve the overall quality of these processes: 1) increase the number of tweets extracted and 2) improve the quality of the labels. Since the beginning of this work, Twitter has improved its APIs to make data extraction easier for researchers. This allows researchers to extract not only more data, but also historical tweets (i.e., tweets in any specific time interval since Twitter's inception). Thus, in future work, we intend to complement the already extracted data with additional tweets from the time interval covered in this thesis as well as previous

dates, hence increasing the overall time interval of the data extracted. This will allow us to assess if, with a longer period of time, the models trained in the experiments conducted in Chapter 5 continue with a stable performance or whether the decline in Figures 5.4, 5.5, 5.8, and 5.9 worsens over time. Additionally and similarly to the study conducted in [112], we would like to measure the impact of new sources on the trained models to understand whether they affect their performance. In addition, given the good results obtained with the BERT and RoBERTa encodings, we also want to investigate how these combined with deep-learning models (such as CNN and LSTM) perform in detecting unreliable and reliable content. Overall, these types of models require large amounts of data. Thus, these experiments are more suitable to be conducted when the previous step of data augmentation is performed. Finally, to improve the constraint of distant labeling, we intend to manually annotate samples of tweets associated with different intervals of time, spread throughout all the data. These will provide additional confidence in the results, as models evaluation will be assessed not only continuously through distant annotation, but also in small samples of manually-annotated data.

Regarding the experiments and the analysis of unreliable accounts, in future work, we intend to implement mechanisms for updating and forgetting accounts based on the proposed metrics. More specifically, we intend to develop tools to update the metrics in each account (i.e., when should an account be removed or updated in our database). Finally, given the new researchers-specific API endpoints, we want to tackle with more precision the time required to extract additional information to complement the proposed metrics and how this information affects the overall computation.

When detecting unreliable accounts, there are still some guidelines that should be considered to better integrate these results into a real-world scenario. In future work, we expect to conduct a more detailed analysis of the solutions presented in Section 4.2 by testing them with new and more recent data. This way, we aim to improve the robustness of the results. We also want to augment the adapted models to include cases where no tweets are provided by the accounts and compare the results obtained with bot detection approaches such as Botometer [243]. Another future step we would like to explore is to approach the problem as a regression task, where instead of discretizing the scores, we predict the reliability of the accounts in its full score spectrum. Although this would require a large investment of human-annotation resources, these experiments would allow us to further distinguish accounts in the same class, providing a severity degree for each account and fine-graining our account reliability detection. Furthermore, this would allow a full unification of the scores of both components studied in Chapter 4 which would be the best solution when developing a system that can be used on a daily basis.

Finally, since the main goal of this work is to apply the detection of unreliable content and accounts to real-world scenarios, we would like to develop an application that combines the

key solutions studied and developed in this thesis: the detection of unreliable accounts and unreliable content in social networks. Ideally, the application would be developed as an add-on for web browsers that allows users to quickly analyze the reliability of a publication or account. In addition, we also intend to explore how the two components can support each other in their respective classification tasks. More specifically, using content detection models to aid in identifying malicious accounts on social networks (e,g., by classifying an account's publication history) and using account detection models to aid in identifying unreliable content (e.g., by using the account metrics/ predictions as features for classifying unreliable content).

6.3 Final Remarks

The dissemination of unreliable information has become more acute with the development of fast-propagation mediums like social networks and, although it has been previously studied, the impact that unreliable information had on the 2016 United States presidential election highlighted the relevance and severity of the problem.

Since 2016, a lot of research has been done towards the mitigation of unreliable content online. Major platforms such as Facebook and Twitter have enforced strategies to prevent the spread of this type of content by either improving mechanisms to detect bots, limiting access to APIs to verified individuals or companies, or working with fact-checking communities to better educate users about the content they are reading/sharing. In the research community several advances were made in the bot and "fake news" detection tasks, as well as extensive and detailed analysis of the dissemination of unreliable information on these platforms.

Nevertheless, four years later, the problem of unreliable content on social media is far from solved. With the global pandemic caused by the new coronavirus, the spread of misinformation has greatly increased due to the uncertainty of the situation. Indeed, unreliable content about Covid-19 has affected millions of lives at a global scale whether due to general disbelief of the virus or, more recently, lack of trust in vaccines. It seemed clear that with the events of 2016, action was needed. Now, given the impact of misinformation, it is more urgent than ever to consider unreliable information on social media as a serious and worrisome problem. Therefore, it is essential to develop systems that can detect this type of content as well as the accounts that disseminate it. To this end, experimental studies must consider real-world constraints as part of their experimental settings. In the work presented in this thesis, we have laid the groundwork for more realistic approaches by either tackling the detection of unreliable accounts considering constraints on the available content, or targeting the detection of unreliable posts over time and with the introduction of new topics. In fact, at the beginning of this work, we could not have anticipated such a large shift in unreliable content topics. In this sense, Covid-19 provided an opportunity to better

explore how the combination of features/models from the revised literature handles such a shift, highlighting the important contribution of this work for similar situations in the future. We recommend that researchers working with unreliable content follow a similar path by approaching the problem in a more time-dependent manner, focusing on the unseen nature of future events. In other words, for content detection we suggest a more chronological approach to the problem by splitting the training and testing of models into two separate time periods. Although we believe that a continuous longitudinal evaluation is the best way to approach the problem, a time gap between the training and test data would result in a more accurate representation of the real world (when compared with the majority of current approaches), and bridge the gap between more experimental and pragmatic solutions. The same principles should be applied to account detection experiments, namely the potential content constraints occurring in a real-world scenario.

However, in the context of this work, a deeper analysis must be carried out. Focusing on the "pragmatic" aspect, which is the predominant "sound byte" of this thesis, one must question how a system based on the presented experiments would affect social networks and whether it would be able to solve the unreliable information problem.

From a perspective where this system is used internally by a social network company such as Facebook or Twitter, this hypothetical application could help flag and remove unreliable content. However, from the user's perspective where the application alerts social media users to the reliability of an account/post, the issue is more complex.

Due to some of the phenomena explored in Chapter 2, such as confirmation bias and echo chambers, the impact of unreliable content detection applications is limited, as in the fact that mainstream news media are discarded as reliable sources by users who are more sympathetic with unreliable content. Moreover, the black-box nature of current approaches is unlikely to be sufficient to convince a user that a particular account/post is unreliable. Therefore, in this scenario, the introduction of explainable AI is necessary to better address skeptical users, and we recommend its integration with this specific task in the future.

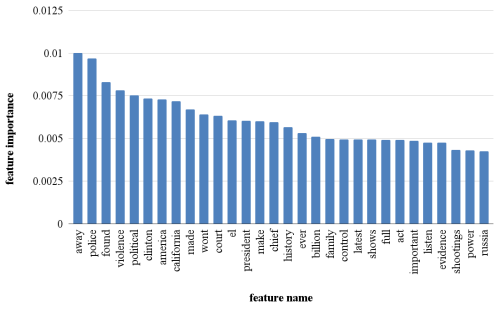
However these systems only help to assess the reliability of accounts and publications (posts), since it is up to users to decide whether they believe the system or not. Hence, news and digital literacy play an important role in the education of an increasingly digitally-aware society, and thus continuous and increasing investments should be made in this area. It is our opinion that combining technological approaches, such as those presented in this work, with high levels of user literacy in digital and news media is the most viable solution to make unreliable information a problem of the past.

Appendix A

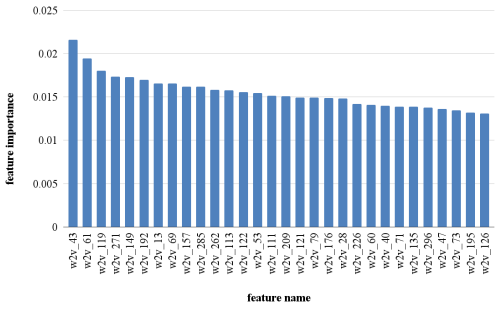
Feature Importance

The following appendix illustrates the feature importance results for the remaining scenarios. Figure A.1 and A.2 refers to the feature importance of the different groups in the 30 and 60 day scenario.

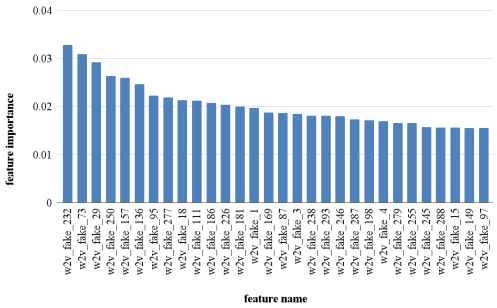
(a) Bag of Words



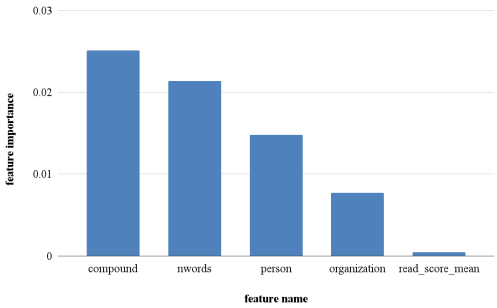
(b) Google Word2Vec



(c) Fake Word2Vec



(d) Context-free features



(e) Lexical Categories (Empath)

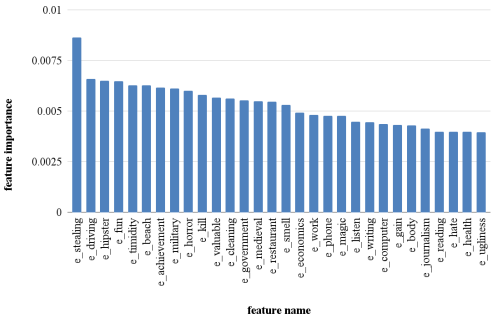
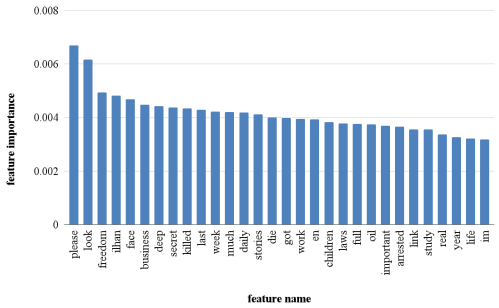
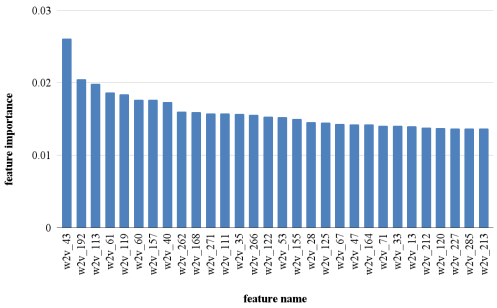


Figure A.1: Importance of the 30 most relevant features from the different feature sets using a 30-day interval. In some scenarios, due to the pre-processing applied (removal of low variance and highly correlated features), the total number of features is inferior to 30.

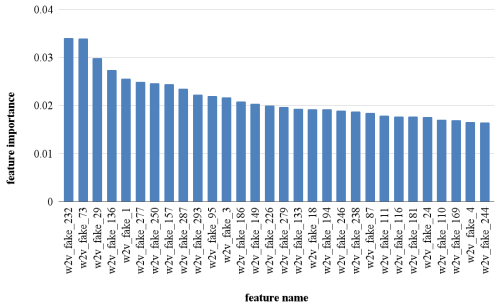
(a) Bag of Words



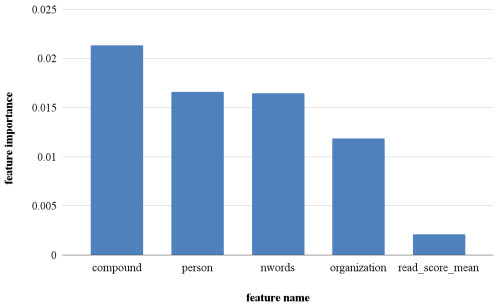
(b) Google Word2Vec



(c) Fake Word2Vec



(d) Context-free features



(e) Lexical Categories (Empath)

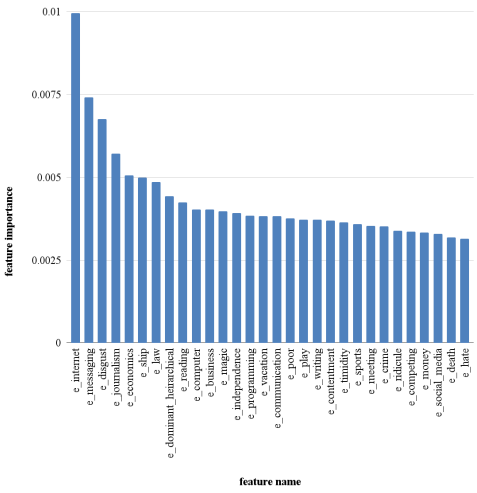


Figure A.2: Importance of the 30 most relevant features from the different feature sets using a 60-day interval. In some scenarios, due to the pre-processing applied (removal of low variance and highly correlated features), the total number of features is inferior to 30.

Appendix B

Models Performance

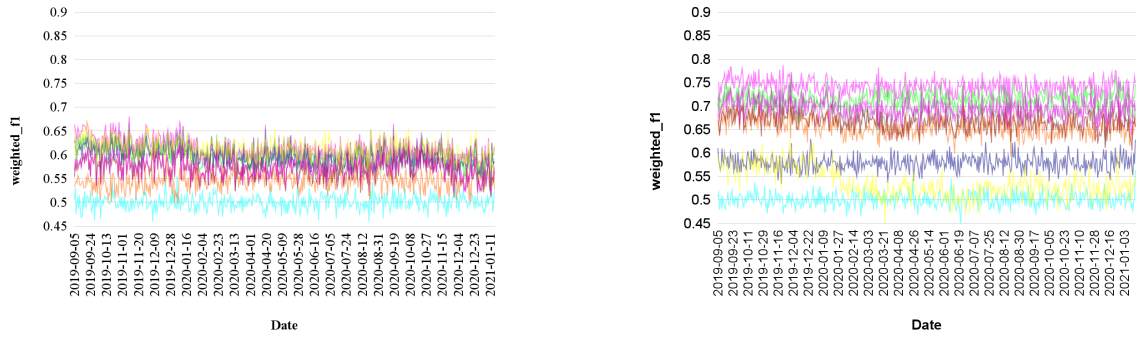
The following appendix presents the models' performance for the remaining scenarios presented in Section 5.3.2. Figure B.1 and B.2 refers to the models performance of the different groups in the 30 and 60 day scenario.

B.1 Models Performance Using BERT and RoBERTa Features

The following appendix presents the models' performance for the remaining scenarios presented in Section 5.4. Figure B.3 and B.4 refers to the models performance of the different groups in the 30 and 60 day scenario.

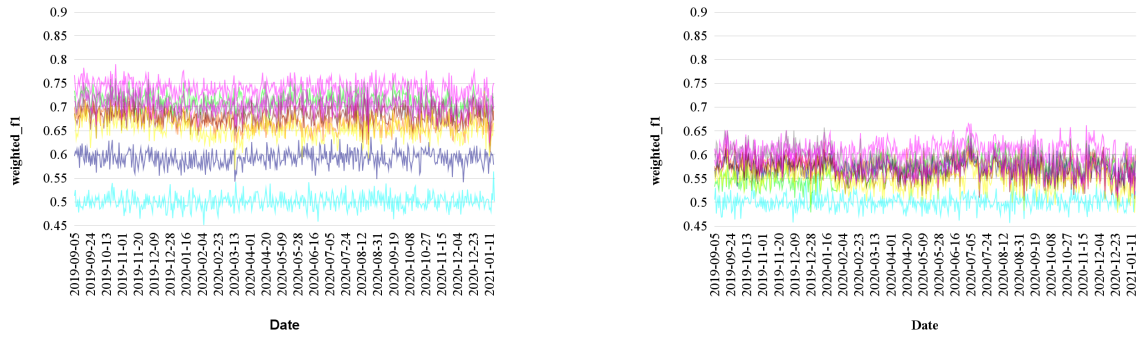
(a) Bag of Words

(b) Google Word2Vec



(c) Fake Word2Vec

(d) Lexical Categories (Empath)



(e) Context-free features

(f) Best 15 features

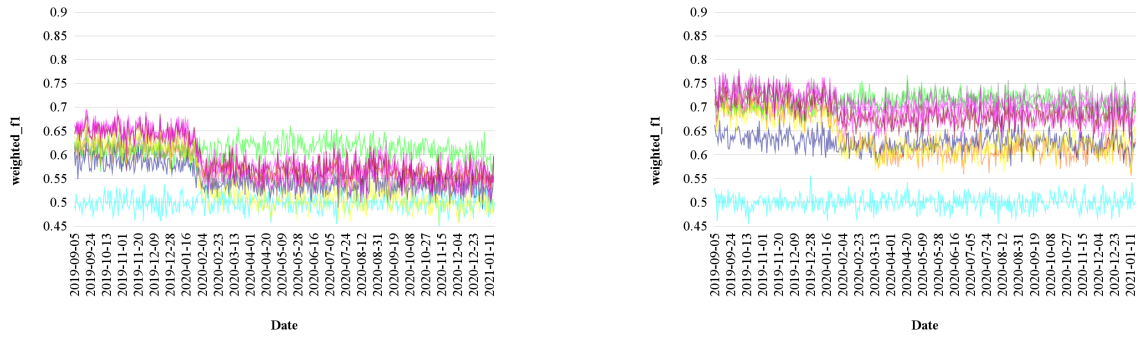


Figure B.1: Performance evaluation (using weighted F1-score) of the different models for each feature set in the 30-day training data

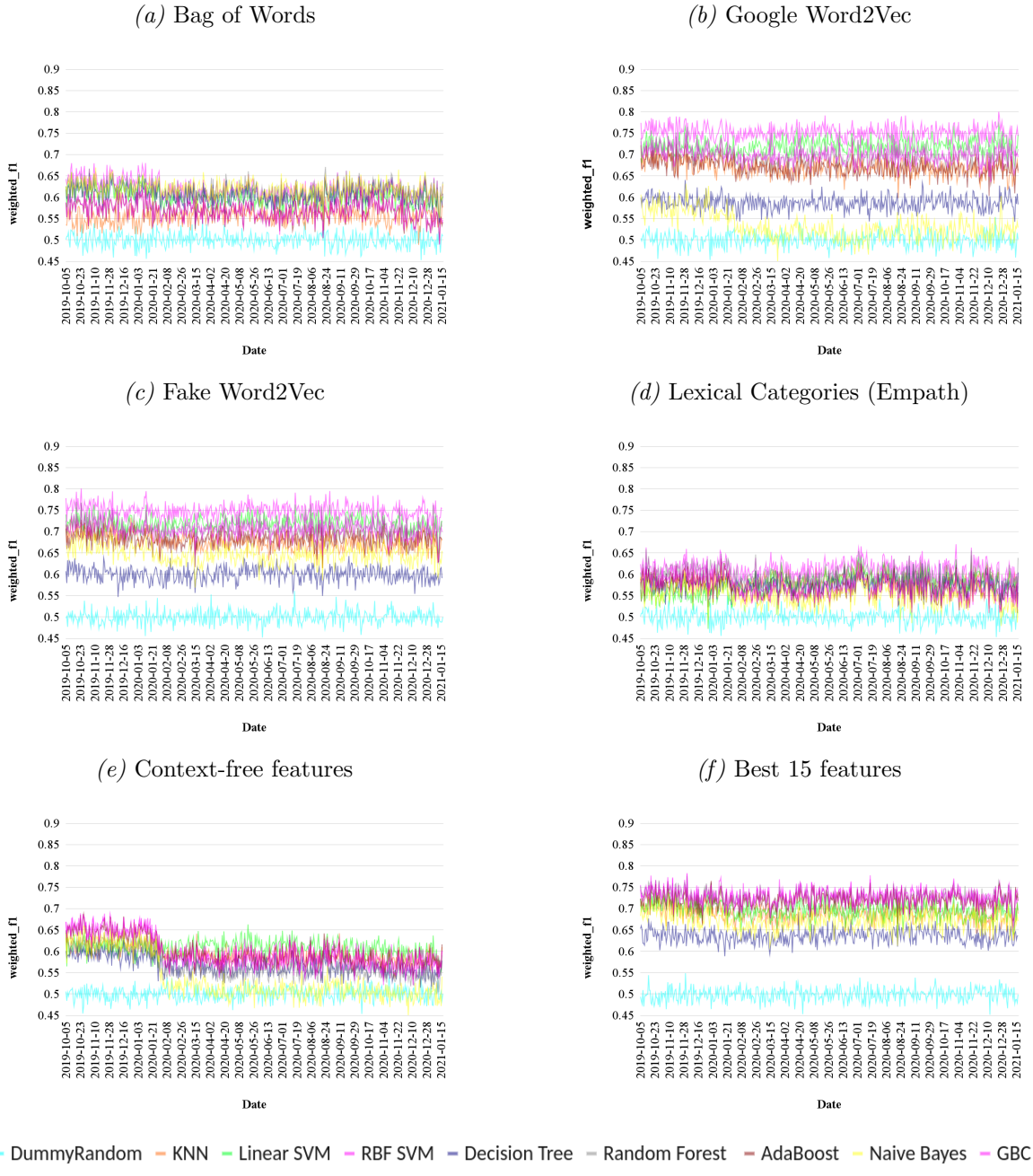
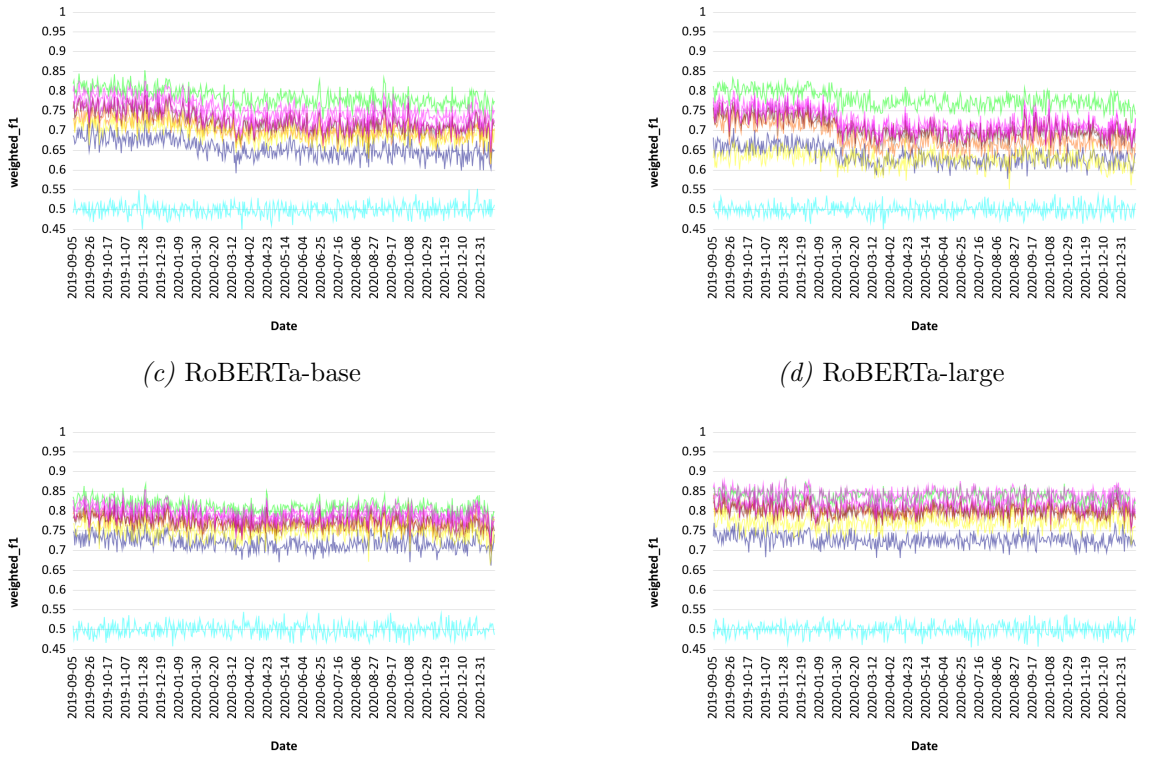


Figure B.2: Performance evaluation (using weighted F1-score) of the different models for each feature set in the 60-day training data

(a) BERT-base

(b) BERT-large



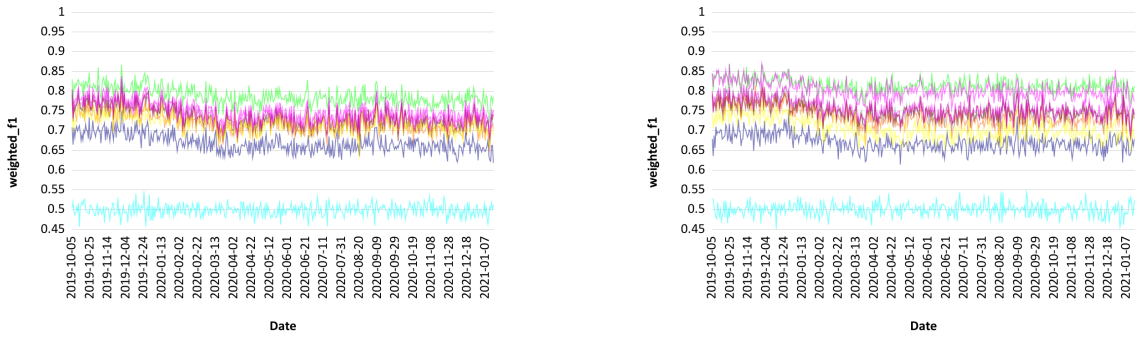
(c) RoBERTa-base

(d) RoBERTa-large

Figure B.3: Performance evaluation (using weighted F1-score) of the different models for each BERT-derived feature set in the 30-day training scenario

(a) BERT-base

(b) BERT-large



(c) RoBERTa-base

(d) RoBERTa-large

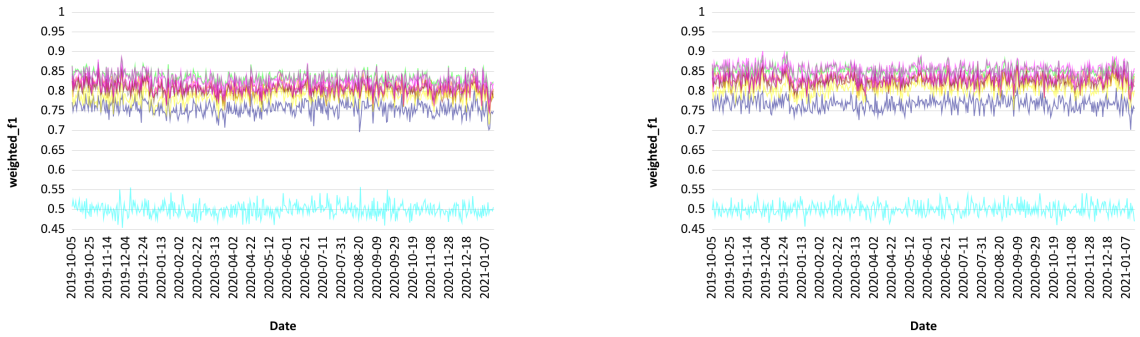


Figure B.4: Performance evaluation (using weighted F1-score) of the different models for each BERT-derived feature set in the 60-day training scenario

References

[1] Ala M. Al-Zoubi, Ja'Far Alqatawna, and Hossam Faris. Spam profile detection in social networks based on public features. *2017 8th International Conference on Information and Communication Systems, ICICS 2017*, pages 130–135, 2017.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May 2017.

[3] Bobby Allyn. Twitter removes thousands of qanon accounts, promises sweeping ban on the conspiracy : Npr. `https://www.npr.org/2020/07/21/894014810/twitter-removes-thousands-of-qanon-accounts-promises-sweeping-ban-on-the-conspir?t=1639162867634&t=1639255416649`, July 2020. (Accessed on 12/11/2021).

[4] N S Altman. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185, 1992.

[5] Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. We used neural networks to detect clickbaits: You won't believe what happened next! In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait, editors, *Advances in Information Retrieval*, pages 541–547, Cham, 2017. Springer International Publishing.

[6] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. *A Model for Identifying Misinformation in Online Social Networks*, volume 9415, pages 473–482. Springer International Publishing, Cham, 2015.

[7] AP. France confirms first three cases of coronavirus in europe. `https://www.cnbc.com/2020/01/24/france-confirms-2-cases-of-virus-from-china-1st-in-europe.html`. (Accessed on 01/26/2022).

[8] AssociatedPress. Brazil's bolsonaro targets minorities on 1st day in office. `https://www.nbcnews.com/feature/nbc-out/brazil-s-bolsonaro-targets-minorities-1st-day-office-n954181`, January 2019. (Accessed on 12/13/2021).

[9] Farzindar Atefeh and Wael Khreich. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[10] Ahmed El Azab, Amira M Idrees, Mahmoud A Mahmoud, and Hesham Hefny. Fake Account Detection in Twitter Based on Minimum Weighted Feature set. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 10(1):13–18, 2016.

[11] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting Factuality of Reporting and Bias of News Media Sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3528–3539, 2018.

[12] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, Brussels, Belgium, 2018.

[13] Ramy Baly, Mitra Mohtarami, James Glass, Lluis Marquez, Alessandro Moschitti, and Preslav Nakov. Integrating Stance Detection and Fact Checking in a Unified Corpus. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3528–3539, 2018.

[14] Donara Barojan. Understanding bots, botnets and trolls. `https://ijnet.org/en/story/understanding-bots-botnets-and-trolls`, 2018. (Accessed on 11/06/2020).

[15] Marco Bastos and Shawn T. Walker. Facebook's data lockdown is a disaster for academic researchers. `https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533`. (Accessed on 01/26/2022).

[16] Marco T. Bastos and Dan Mercea. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, 37(1):38–54, 2019.

[17] BBCNews. 'cambridge analytica planted fake news' - bbc news. `https://www.bbc.com/news/av/world-43472347`, March 2018. (Accessed on 12/11/2021).

[18] BBCNews. Venezuela president maduro survives 'drone assassination attempt' - bbc news. `https://www.bbc.com/news/world-latin-america-45073385`, August 2018. (Accessed on 12/11/2021).

[19] BBCNews. Syria war: 'is suicide bomber' kills us troops in manbij - bbc news. `https://www.bbc.com/news/world-middle-east-46892118`, January 2019. (Accessed on 12/13/2021).

[20] F Benevenuto, G Magno, T Rodrigues, and V Almeida. Detecting spammers on twitter. *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 6:12, 2010.

[21] David M. Beskow and Kathleen M. Carley. Bot conversations are different: Leveraging network metrics for bot detection in Twitter. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pages 825–832, 2018.

[22] David M. Beskow and Kathleen M. Carley. *You Are Known by Your Friends: Leveraging Network Metrics for Bot Detection in Twitter.* Springer International Publishing, 2020.

[23] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE*, 10(2):1–17, 2015.

[24] Carolina Bigonha, Thiago N.C. Cardoso, Mirella M. Moro, Marcos A. Gonçalves, and Virgílio A.F. Almeida. Sentiment-based influence detection on Twitter. *Journal of the Brazilian Computer Society*, 18(3):169–183, 2012.

[25] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly, Beijing, 2009.

[26] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, mar 2003.

[27] Laura Bogart and Sheryl Thorburn. Are HIV/AIDS conspiracy beliefs a barrier to HIV prevention among African Americans? *Journal of acquired immune deficiency syndromes (1999)*, 38:213–218, 03 2005.

[28] Thomas Boghardt. Soviet Bloc Intelligence and Its AIDS Disinformation Campaign, 2009.

[29] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[30] Alexandre Bovet and Hernan A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10:1–23, 2018.

[31] Alexandre Bovet and Hernan A Hernán A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1):1–14, 1 2019.

[32] Emma Brown. California professor christine blasey ford, writer of confidential brett kavanaugh letter, speaks out about sexual assault allegation - the washington post. `https://www.washingtonpost.com/investigations/california-professor-writer-of-confidential-brett-kavanaugh-letter-speaks-out-about-her-`

allegation-of-sexual-assault/2018/09/16/46982194-b846-11e8-94eb-3bd52dfe917b_story.html, September 2018. (Accessed on 12/11/2021).

[33] Jakab Buda and Flora Bolonyai. An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter Notebook for PAN at CLEF 2020, September 2020.

[34] Peter Burger, Soeradj Kanhai, Alexander Pleijter, and Suzan Verberne. The reach of commercially motivated junk news on Facebook. *PLoS ONE*, 14(8):1–15, 2019.

[35] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach — cambridge analytica — the guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, March 2018. (Accessed on 12/11/2021).

[36] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T. Marques-Neto, and Virgílio Almeida. Characterizing attention cascades in whatsapp groups. In *WebSci 2019 - Proceedings of the 11th ACM Conference on Web Science*, 2019.

[37] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[38] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. *Proceedings of the 20th international conference on World wide web*, 2011.

[39] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, number May 2014 in WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[40] CDC. Cdc museum covid-19 timeline. https://www.cdc.gov/museum/timeline/covid19.html. (Accessed on 01/26/2022).

[41] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

[42] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as "false news". *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19, 2015.

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

[44] Soon Ae Chun, Richard Holowczak, Kannan Neten Dharan, Ruoyu Wang, Soumaydeep Basu, and James Geller. Detecting political bias trolls in Twitter data. *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, pages 334–342, 2019.

[45] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6):1–13, 2015.

[46] Matteo Cinelli, Stefano Cresci, Alessandro Galeazzi, Walter Quattrociocchi, and Maurizio Tesconi. The limited reach of fake news on Twitter during 2019 European elections. *PLoS ONE*, 15(6):1–13, 2020.

[47] 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. *Scientific Reports*, 10(1):1–10, 2020.

[48] CNN. What we know about the Boston bombing and its aftermath. https://edition.cnn.com/2013/04/18/us/boston-marathon-things-we-know, 2013. Acessed: 2018-06-12.

[49] Kate Conger. Twitter removes chinese disinformation campaign. `https://www.nytimes.com/2020/06/11/technology/twitter-chinese-misinformation.html`, June 2020. Acessed: 2020-07-07.

[50] David Conn. How the Sun's 'truth' about Hillsborough unravelled. https://www.theguardian.com/football/2016/apr/26/how-the-suns-truth-about-hillsborough-unravelled, 2016. Acessed: 2018-06-07.

[51] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. In *Machine Learning*, volume 20, pages 273–297. Kluwer Academic Publishers, September 1995.

[52] Jin Dan and Teng Jieqi. Study of Bot detection on Sina-Weibo based on machine learning. *14th International Conference on Services Systems and Services Management, ICSSSM 2017 - Proceedings*, 2017.

[53] Anh Dang, Abidalrahman Moh', Aminul Islam, and Evangelos Milios. Early Detection of Rumor Veracity in Social Media. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 6:2355–2364, 2019.

[54] Anh Dang, Michael Smit, Abidalrahman Moh'D, Rosane Minghim, and Evangelos Milios. Toward understanding how users respond to rumours in social media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, pages 777–784, 2016.

[55] Kheir Eddine Daouadi, Rim Zghal Rebaï, and Ikram Amous. Bot detection on online social networks using deep forest. *Advances in Intelligent Systems and Computing*, 985:307–315, 2019.

[56] Datareportal. Social media users by platform. `https://datareportal.com/social-media-users`, 2019.

[57] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. BotOrNot: A System to Evaluate Social Bots. In *WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web*, page 273–274. ACM, ACM, 2016.

[58] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web*, 13(2), 2019.

[59] Leon Derczynski, Torben Oskar Albert-Lindqvist, Marius Venø Bendsen, Nanna Inie, Viktor Due Pedersen, and Jens Egholm Pedersen. Misinformation on Twitter During the Danish National Election: A Case Study. In *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019*, 2019.

[60] Prateek Dewan and Ponnurangam Kumaraguru. Towards automatic real time identification of malicious posts on Facebook. *2015 13th Annual Conference on Privacy, Security and Trust, PST 2015*, pages 85–92, 2015.

[61] John P Dickerson, Vadim Kagan, and V S Subrahmanian. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 620–627, 2014.

[62] Júlio Cesar dos Reis, Fabrıcio Benevenuto, Pedro Olmo Stancioli Vaz de Melo, Raquel O Prates, Haewoon Kwak, and Jisun An. Breaking the News: First Impressions Matter on Online News. *CoRR*, abs/1503.0, 2015.

[63] Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. Can Rumour Stance Alone Predict Veracity? *Proceedings of the 27th International Conference on Computational Linguistics*, August 2018.

[64] Jillian D'Onfro. Twitter responds to misinformation after youtube shooting. `https://www.cnbc.com/2018/04/05/twitter-responds-to-misinformation-after-youtube-shooting.html`, April 2018. (Accessed on 12/11/2021).

[65] Adam Ellick and Adam Westbrook. Operation Infektion - Russian Disinformation: From Cold War to Kanye. `https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html`, 2018.

[66] Philip Elliott. John mccain dies at 81 after brain cancer battle — time. `https://time.com/5179302/john-mccain-dies/`, August 2018. (Accessed on 12/11/2021).

[67] Buket Erşahin, Özlem Aktaş, Deniz Kilmç, and Ceyhun Akyol. Twitter fake account detection. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, pages 388–392, 2017.

[68] Facebook. Restricting data access and protecting people's information on facebook — facebook for business. `https://www.facebook.com/business/news/restricting-data-access-and-protecting-peoples-information-on-facebook`, April 2018. (Accessed on 12/11/2021).

[69] Facebook. New analytics api for researchers studying facebook page data - facebook research. `https://research.fb.com/blog/2021/03/new-analytics-api-for-researchers-studying-facebook-page-data/`, March 2021. (Accessed on 12/11/2021).

[70] FakeNewsChallenge.org. Stance Detection dataset for FNC-1. https://github.com/FakeNewsChallenge/fnc-1, 2017. Acessed: 2018-04-12.

[71] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In *34th International Conference on Machine Learning, ICML 2017*, 2017.

[72] Johan Farkas, Jannick Schou, and Christina Neumayer. Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media and Society*, 2017.

[73] Ethan Fast, Binbin Chen, and Michael Bernstein. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA, 2016. Association for Computing Machinery.

[74] Emilio Ferrara. What Types of Covid-19 Conspiracies Are Populated By Twitter Bots? *arXiv*, 2020.

[75] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 6 2016.

[76] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168. Association for Computational Linguistics, 2016.

[77] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, 2017.

[78] Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[79] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[80] Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. ExFaKT. *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 87–95, January 2019.

[81] Jaynil Gaglani, Yash Gandhi, Shubham Gogate, and Aparna Halbe. Unsupervised WhatsApp Fake News Detection using Semantic Search. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 285–289, 2020.

[82] Joao Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4):258–264, 2010.

[83] Lazaro Gamio and Callum Borchers. A visual history of donald trump dominating the news cycle, 2016. https://www.washingtonpost.com/graphics/politics/donald-trump-vs-hillary-clinton-in-media/.

[84] Shirin Ghaffary. Twitter is verifying people again. here's how you can get a blue checkmark. - vox. `https://www.vox.com/22444961/twitter-verification-process-verified-blue-checkmark-jack-dorsey`, May 2021. (Accessed on 12/11/2021).

[85] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. Classification of Twitter Accounts into Automated Agents and Human Users. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17*, pages 489–496, 2017.

[86] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran. How WhatsApp Leads Mobs to Murder in India. `https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html`, 2018. Online; posted on 18-Jul-2018.

[87] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jeannine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenye Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21, October 2018.

[88] Genevieve Gorrell, Ian Roberts, Mark A. Greenwood, Mehmet E. Bakir, Benedetta Iavarone, and Kalina Bontcheva. *Quantifying media influence and partisan attention on twitter during the UK EU referendum*, volume 11185 LNCS. Springer International Publishing, 2018.

[89] Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(1), 2017.

[90] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, 2019.

[91] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles, 2020.

[92] Stefano Guarino, Noemi Trino, Alessandro Celestini, Alessandro Chessa, and Gianni Riotta. Characterizing networks of propaganda on twitter: a case study. *Applied Network Science*, 5(1), 2020.

[93] Nuno Guimaraes, Alvaro Figueira, and Luis Torgo. Contributions to the Detection of Unreliable Twitter Accounts through Analysis of Content and Behaviour. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, September 18-20, 2018.*, pages 90–99, 2018.

[94] Nuno Guimaraes, Alvaro Figueira, and Luis Torgo. Analysis and Detection of Unreliable Users in Twitter: Two Case Studies. In Ana Fred, Ana Salgado, David Aveiro, Jan Dietz, Jorge Bernardino, and Joaquim Filipe, editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 1222 CCIS

of *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 50–73, Cham, June 2020. Springer International Publishing.

[95] Aditi Gupta. Twitter Explodes with Activity in Mumbai Blasts! A Lifeline or an Unmonitored Daemon in the Lurking? *Precog.Iiitd.Edu.in*, pages 1–17, September 2011.

[96] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on twitter. *eCrime Researchers Summit, eCrime*, 2013.

[97] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 729–736, 2013.

[98] M Gupta, J Gao, C Zhai, and J Han. Predicting Future Popularity Trend of Events in Microblogging Platforms. In Andrew Grove, editor, *ASIS&T 75th Annual Meeting*, 2012.

[99] Supraja Gurajala, Joshua S. White, Brian Hudson, and Jeanna N. Matthews. Fake Twitter Accounts: Profile Characteristics Obtained Using an Activity-based Pattern Detection Approach. *Proceedings of the 2015 International Conference on Social Media & Society*, pages 9:1–9:7, 2015.

[100] B. Hajian and T. White. Modelling influence in a social network: Metrics and evaluation. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 497–500, 2011.

[101] Sardar Hamidian and Mona T Diab. Rumor Detection and Classification for Twitter Data. *Proceedings of SOTICS 2015 : The Fifth International Conference on Social Media Technologies, Communication, and Informatics*, pages 71–77, 2015.

[102] David J. Hand and Keming Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review / Revue Internationale de Statistique*, 69(3):385–398, 2001.

[103] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A Survey on Stance Detection for Mis- and Disinformation Identification. *preprint*, 2021.

[104] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on Twitter. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pages 274–277, 2018.

[105] Alex Hern. Google acts against fake news on search engine. `https://www.theguardian.com/technology/2017/apr/25/google-launches-major-offensive-against-fake-news`, 2017. Accessed: 2018-04-13.

[106] Alex Hern. Facebook, apple, youtube and spotify ban infowars' alex jones — apple — the guardian. `https://www.theguardian.com/technology/2018/aug/06/apple-removes-podcasts-infowars-alex-jones`, August 2018. (Accessed on 12/11/2021).

[107] Alex Hern. New Facebook controls aim to regulate political ads and fight fake news. `https://www.theguardian.com/technology/2018/apr/06/facebook-launches-controls-regulate-ads-publishers`, 2018. Accessed: 2018-04-13.

[108] Alex Hern. Twitter to remove harmful fake news about coronavirus, 2020. https://www.theguardian.com/world/2020/mar/19/twitter-to-remove-harmful-fake-news-about-coronavirus.

[109] Tucker Higgins. Trump picks brett kavanaugh for the supreme court. `https://www.cnbc.com/2018/07/05/trump-picks-brett-kavanaugh-for-supreme-court.html`, July 2018. (Accessed on 12/11/2021).

[110] Nathan Hodge and Darya Tarasova. Chechnya: deaths and detentions in 'new wave of persecution,' say lgbt activists - cnn. `https://edition.cnn.com/2019/01/14/europe/russian-lgbt-activists-crackdown-chechnya-intl/index.html`, January 2019. (Accessed on 12/13/2021).

[111] Benjamin D. Horne, Sara Khedr, and Sibel Adalı. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *arXiv*, 2018.

[112] Benjamin D. Horne, Jeppe NØrregaard, and Sibel Adali. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology*, 11(1):1–23, 2019.

[113] Philip N. Howard and Bence Kollanyi. Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. *SSRN Electronic Journal*, 2017.

[114] Xia Hu, Jiliang Tang, and Huan Liu. Online social spammer detection. *Proceedings of the National Conference on Artificial Intelligence*, 1:59–65, 2014.

[115] Muhammad Nihal Hussain, Serpil Tokdemir, Nitin Agarwal, and Samer Al-Khateeb. Analyzing disinformation and crowd manipulation tactics on youtube. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pages 1092–1095, 2018.

[116] Clayton J Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H Oh, editors, *ICWSM*. The AAAI Press, 2014.

[117] Samidha Jain and Kapil Kashyap. Whatsapp: Indians most active on whatsapp with 390.1 million monthly active users in 2020. `https://www.forbesindia.com/article/news-by-numbers/indians-most-active-on-whatsapp-with-3901-million-monthly-active-users-in-2020/70059/1`. (Accessed on 01/21/2022).

[118] Christian Janze and Marten Risius. Automatic Detection of Fake News on Social Media Platforms. *PACIS 2017 Proceedings*, 2017.

[119] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, pages 795–816, 2017.

[120] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10354 LNCS(October):14–24, 2017.

[121] Zhiwei Jin, Juan Cao, Yu Gang Jiang, and Yongdong Zhang. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2015-Janua(January):230–239, 2015.

[122] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia*, 19(3):598–608, 2017.

[123] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788, 2021.

[124] William Ogilvy Kermack and A G McKendrick. A contribution to the mathematical theory of epidemics. *The American Mathematical Monthly*, 45(7):446, 1938.

[125] Eugene Kiely. MultiFC : A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. *arxiv*, September 2019.

[126] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, David Corney, Payam Adineh, Benno Stein, and Martin Potthast. Data for PAN at SemEval 2019 Task 4: Hyperpartisan News Detection, 11 2018.

[127] Vanessa L Kitzie, Ehsan Mohammadi, and Amir Karami. "Life never matters in the DEMOCRATS MIND": Examining strategies of retweeted social bots during a mass shooting event. *Proceedings of the Association for Information Science and Technology*, 55(1):254–263, 2018.

[128] Rebecca Klein. An Army Of Sophisticated Bots Is Influencing The Debate Around Education. `https://www.huffingtonpost.com/entry/common-core-debate-bots{_}us{_}58bc8bf3e4b0d2821b4ee059`, 2017. Acessed: 2018-05-07.

[129] Elizabeth A Klonoff and Hope Landrine. Do Blacks Believe That HIV/AIDS Is a Government Conspiracy against Them? *Preventive Medicine*, 28(5):451–457, 1999.

[130] Saranya Knshnan and Min Chen. Identifying tweets with fake news. *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 67:460–464, 2018.

[131] Bence Kollanyi, Philip N. Howard, and Samuel C. Woolley. Bots and Automation over Twitter during the First U.S. Election. *Data Memo*, pages 1–5, 2016.

[132] Paweł Ksieniewicz, Paweł Zyblewski, Michał Choraś, Rafał Kozik, Agata Giełczyk, and Michał Woźniak. Fake News Detection from Data Streams. *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[133] Ema Kušen and Mark Strembeck. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, 5:37–50, 2018.

[134] Issie Lapowsky. Cambridge analytica execs caught discussing extortion and fake news — wired. `https://www.wired.com/story/cambridge-analytica-execs-caught-discussing-extortion-and-fake-news/`, March 2018. (Accessed on 12/11/2021).

[135] Georgina Lee. FactCheck: the broken Brexit promises, half-truths and dodgy predictions from all sides – Channel 4 News. `https://www.channel4.com/news/factcheck/factcheck-the-broken-brexit-promises-half-truths-and-dodgy-predictions-from-all-sides`, 1 2020.

[136] Chenliang Li, Aixin Sun, and A Datta. Twevent: Segment-based Event Detection from Tweets. *Cikm*, pages 155–164, 2012.

[137] Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, and Sameena Shah. User Behaviors in Newsworthy Rumors : A Case Study of Twitter. *Proceedings of the Tenth International AAAI Conference*, pages 627–630, 2016.

[138] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. Real-time and cost-effective limitation of misinformation propagation. *Proceedings - IEEE International Conference on Mobile Data Management*, 2016-July:158–163, 2016.

[139] Chao Liu, Xinghua Wu, Min Yu, Gang Li, Jianguo Jiang, Weiqing Huang, and Xiang Lu. *A Two-Stage Model Based on BERT for Short Fake News Detection*, volume 11776 LNAI. Springer International Publishing, 2019.

[140] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arxiv*, 2019.

[141] Clare Llewellyn, Laura Cram, Robin L. Hill, and Adrian Favero. For Whom the Bell Trolls: Shifting Troll Behaviour in the Twitter Brexit Debate. *Journal of Common Market Studies*, 57(5):1148–1164, 2019.

[142] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson. Measuring the Impact of Exposure to COVID-19 Vaccine Misinformation on Vaccine Intent in the UK and US. *Nature Human Behaviour*, 2020.

[143] Eugène Loos and Jordy Nijenhuis. *Consuming Fake News: A Matter of Age? The Perception of Political Fake News Stories in Facebook Ads*, volume 12209 LNCS. Springer International Publishing, 2020.

[144] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect Rumor and Stance Jointly by Neural Multi-task Learning. *Companion Proceedings of the The Web Conference 2018*, 1(Long Papers):585–593, 2018.

[145] Josh Margolin and Catherine Thorbecke. Twitter removes account of white nationalist group posing as antifa online. `https://abcnews.go.com/US/twitter-removes-account-white-nationalist-group-posing-antifa/story?id=71024345`, June 2020. Acessed: 2020-07-07.

[146] Julian Marx, Felix Brünker, and Eric Hochstrate. ‘ Conspiracy Machines ’ - The Role of Social Bots during the COVID-19 ‘ Infodemic ’. In *Proceedings of ACIS 2020*, pages 1–8, 2020.

[147] MediaBias. Media bias/fact check -the most comprehensive media bias resources. `https://mediabiasfactcheck.com/`. Acessed: 2018-05-03.

[148] Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, and Jussara Almeida. WhatsApp Monitor : A Fact-Checking System for WhatsApp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(April):6–8, Jul. 2019.

[149] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter Under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics*, number January in SOMA ’10, pages 71–79, New York, NY, USA, 2010. ACM.

[150] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3, 2013.

[151] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Matsa. Political Polarization & Media Habits. `http://www.journalism.org/2014/10/21/political-polarization-media-habits/`, 2014.

[152] Amy Mitchell, Mark Jurkowitz, J Oliphant, and Elisa Shearer. How Americans Navigated the News in 2020: A Tumultuous Year in Review. *Pew Research Center*, 2021.

[153] Tanushree Mitra and Eric Gilbert. CREDBANK: A Large-scale Social Media Corpus With Associated Credibility Annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 80, pages 787–788, 2015.

[154] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[155] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[156] Damian Mrowca and Elias Wang. Stance Detection for Fake News Identification, 2017.

[157] Robert S Mueller. Report on the Investigation into Russian Interference in the 2016 Presidential Election, 2019.

[158] Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking approaches for microblog search. *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, 1(Section III):153–157, 2010.

[159] I. Natividad. Covid-19 and the media: The role of journalism in a global pandemic, 2020. https://news.berkeley.edu/2020/05/06/covid-19-and-the-media-the-role-of-journalism-in-a-global-pandemic/.

[160] Indira Neill Hoch. Russian Internet Research Agency Disinformation Activities on Tumblr: Identity, Privacy, and Ambivalence. *Social Media and Society*, 6(4), 2020.

[161] Pontus Nordberg, Joakim Kävrestad, and Marcus Nohlberg. Automatic detection of fake news. *CEUR Workshop Proceedings*, 2789(September):168–179, 2020.

[162] Richard Norton-Taylor. Zinoviev letter was dirty trick by MI6. https://www.theguardian.com/politics/1999/feb/04/uk.politicalnews6, 1999. Acessed: 2018-06-07.

[163] OpenSources. OpenSources - Professionally curated lists of online sources, available free for public use. https://github.com/OpenSourcesGroup/opensources, 2018. Acessed: 2018-05-03.

[164] Michael Orlov and Marina Litvak. Using Behavior and Text Analysis to Detect Propagandists and Misinformers on Twitter. In Juan Antonio Lossio-Ventura, Denisse Muñante, and Hugo Alatrista-Salas, editors, *Information Management and Big Data*, volume 898, pages 67–74, Cham, February 2019. Springer International Publishing.

[165] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, pages 45–54, 2011.

[166] PAN. Hyperpartisan News Detection. `"https://pan.webis.de/semeval19/semeval19-web/"`. [Accessed: 2019-03-14].

[167] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu Liu. *Content Based Fake News Detection Using Knowledge Graphs*, volume 11136. Springer International Publishing, 2018.

[168] Gordon Pennycook and David G. Rand. Crowdsourcing Judgments of News Source Quality. *SSRN Electronic Journal*, pages 1–26, 2018.

[169] Micael Pereira. Fake news no twitter sobre vírus quintuplicam num mês. `https://expresso.pt/sociedade/2020-03-28-Fake-news-no-Twitter-sobre-virus-quintuplicam-num-mes`, March 2020. (Accessed on 12/11/2021).

[170] Micael Pereira. A verdade que nós andamos a esconder sobre a covid. `https://expresso.pt/sociedade/2020-06-07-A-verdade-que-nos-andamos-a-esconder-sobre-a-covid`, June 2020. (Accessed on 12/11/2021).

[171] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. *arxiv*, page 3291382, 2017.

[172] Stephen Pfohl, Oskar Triebe, and Ferdinand Legros. Stance Detection for the Fake News Challenge with Attention and Conditional Encoding. *preprint*, pages 1–14, 2016.

[173] Juan Pizarro. Using N-grams to detect Fake News Spreaders on Twitter. *Cappellato, L., Eickhoff, C., Ferro, N., N´ev´eol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers*, pages 22–25, 2020.

[174] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. *International Conference on Information and Knowledge Management, Proceedings*, 24-28-Octo:2173–2178, 2016.

[175] Ben Popken. Twitter deleted 200,000 Russian troll tweets. Read them here., 2018. [Online; accessed 13-March-2019].

[176] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[177] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. BuzzFeed-Webis Fake News Corpus 2016, 2 2018.

[178] S. Primario, D. Borrelli, G. Zollo, L. Iandoli, and C. Lipizzi. Measuring polarization in Twitter enabled in online political conversation: The case of 2016 US Presidential election. *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, 2017-Janua:607–613, 2017.

[179] Savyan Pv and S. Mary Saira Bhanu. UbCadet: detection of compromised accounts in twitter based on user behavioural profiling. *Multimedia Tools and Applications*, 79(27-28):19349–19385, 2020.

[180] Ethar Qawasmeh, Mais Tawalbeh, and Malak Abdullah. Automatic Identification of Fake News Using Deep Learning. In *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, 2019.

[181] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo Chambers on Facebook. *ssrn*, 2016.

[182] Jonathan D. Quick and Heidi Larson. The vaccine-autism myth started 20 years ago. here's why it endures today — time. `https://time.com/5175704/andrew-wakefield-vaccine-autism/`, February 2018. (Accessed on 12/13/2021).

[183] J R Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[184] Neel Rakholia and Shruti Bhargava. Is it true? – Deep Learning for Stance Detection in News. *preprint*, 2017.

[185] Francisco Rangel, Paolo Rosso, Bilal Ghanem, and Anastasia Giachanou. PAN @ CLEF 2020 - Profiling Fake News Spreaders on Twitter. `https://pan.webis.de/clef20/pan20-web/author-profiling.html`, 2020.

[186] Adrian Rauchfleisch and Jonas Kaiser. The False Positive Problem of Automatic Bot Detection in Social Science Research. *SSRN Electronic Journal*, 15(October), 2020.

[187] Raymond S. Nickerson. Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[188] Gerasimos Razis and Ioannis Anagnostopoulos. Influencetracker: Rating the impact of a twitter account. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *Artificial Intelligence Applications and Innovations*, pages 184–195, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[189] Pew Research. States of News Media. https://www.journalism.org/2019/06/25/archived-state-of-the-news-media-reports/, 2019.

[190] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabricio Benevenuto. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference*, WWW '19, page 818–828, New York, NY, USA, 2019. Association for Computing Machinery.

[191] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on Twitter: A survey. *Information Processing and Management*, 52(5):949–975, 2016.

[192] Megan Risdal. Getting Real about Fake News. https://www.kaggle.com/mrisdal/fake-news, 2016. Acessed: 2019-03-14.

[193] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, 2015.

[194] Twitter Safety. An update following the riots in washington, dc. https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--, January 2021. (Accessed on 12/11/2021).

[195] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[196] Giovanni C. Santia, Munif Ishad Mujib, and Jake Ryland Williams. Detecting social bots on facebook in an information veracity context. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, pages 463–472, 2019.

[197] Giovanni C. Santia and Jake Ryland Williams. Buzz face: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the 12th International AAAI Conference on Web and Social Media, ICWSM 2018*, pages 531–540, August 2018.

[198] Rüdiger Schmitt-Beck. Bandwagon Effect. *The International Encyclopedia of Political Communication*, pages 1–5, 2015.

[199] Vinay Setty and Erlend Rekve. Truth be Told: Fake News Detection Using User Reactions on Reddit. *International Conference on Information and Knowledge Management, Proceedings*, pages 3325–3328, 2020.

[200] Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, 3:279–317, 2020.

[201] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 745–750, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[202] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kaicheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 2017.

[203] Chengcheng Shao, Pik Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *PLoS ONE*, 13(4):1–23, 2018.

[204] Elisa Shearer and Katerina Eva Matsa. News Use Across Social Media Platforms 2018, 2018.

[205] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Finding Streams in Knowledge Graphs to Support Fact Checking. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2017-Novem:859–864, 2017.

[206] John Merriman Sholar, Shahil Chopra, and Saachi Jain. Towards Automatic Identification of Fake News : Headline-Article Stance Detection with LSTM Attention Models, 2017.

[207] Kai Shu, H. Russell Bernard, and Huan Liu. Studying Fake News via Network Analysis: Detection and Mitigation. *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pages 43–65, 2019.

[208] Kai Shu and Huan Liu. Detecting Fake News on Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2019.

[209] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188, 2020.

[210] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. In

Proceedings of the International AAAI Conference on Web and Social Media, March 2019.

[211] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. *Sigkdd*, 19(1):22–36, 2017.

[212] Kai Shu, Suhang Wang, and Huan Liu. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. Institute of Electrical and Electronics Engineers Inc., 6 2018.

[213] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 312–320, 2019.

[214] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. Tweet stance detection using an attention based neural ensemble model. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1868–1873, 2019.

[215] Craig Silverman and Alexander Lawrence. How Teens In The Balkans Are Duping Trump Supporters With Fake News, 2016.

[216] Craig Silverman, Jane Lytvynenko, Lam Vo, and Jeremy Singer-Vine. Inside The Partisan Fight For Your News Feed, 2017. [Online; accessed 13-March-2019].

[217] Aaron Souppouris. Clickbait, fake news and the power of feeling. `https://www.engadget.com/2016/11/21/clickbait-fake-news-and-the-power-of-feeling/`, 2016. Accessed: 2018-05-07.

[218] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings*, 2014.

[219] Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019. `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`, 2019.

[220] Statista. Most used social media 2020 — Statista. `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, 2021.

[221] Ian Stewart. Jair bolsonaro, right-wing populist, sworn in as president of brazil : Npr. `https://www.npr.org/2019/01/01/681429911/right-wing-populist-jair-bolsonaro-sworn-in-as-president-of-brazil`, January 2019. (Accessed on 12/13/2021).

[222] Maciej Szpakowski. Fake News Corpus. https://github.com/several27/FakeNewsCorpus, 2018.

[223] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some Like it Hoax: Automated Fake News Detection in Social Networks. *ArXiv e-prints*, pages 1–12, April 2017.

[224] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 Earthquake: A twitter case study. *PLoS ONE*, 10(4):1–18, 2015.

[225] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks. *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 977–982, 2015.

[226] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[227] L L C The Self Agency. B.S. Detector - A browser extension that alerts users to unreliable news sources. `http://bsdetector.tech/`, 2016. Accessed: 2018-06-18.

[228] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zian Wang. Trusting Tweets : The Fukushima Disaster and Information Source Credibility on Twitter. *Iscram*, pages 1–10, April 2012.

[229] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. *arXiv*, pages 809–819, 2018.

[230] Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. *Early Detection of Rumours on Twitter via Stance Transfer Learning*, volume 12035 LNCS. Springer International Publishing, 2020.

[231] Craig Timberg and Elizabeth Dwoskin. Exclusive: Twitter is suspending millions of bots and fake accounts every day to fight disinformation - the washington post. `https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/`, July 2018. (Accessed on 12/11/2021).

[232] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE*, 13(9):1–21, 2018.

[233] Nicolas Turenne. The rumour spectrum. *PLoS ONE*, 13(1):1–27, 2018.

[234] Twitter. Twitter verification faq — twitter help. https://help.twitter.com/en/managing-your-account/twitter-verified-accounts. (Accessed on 12/11/2021).

[235] Twitter. Twitter - Search API. "https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets", 2018. [Accessed: 2018-05-07].

[236] Twitter. Twitter trends faqs. "https://help.twitter.com/en/using-twitter/twitter-trending-faqs", 2018. [Accessed: 2019-01-16].

[237] Twitter. Twitter Verified. https://twitter.com/verified, 2018. Acessed: 2018-05-17.

[238] Twitter. Twitter - filter realtime tweets. "https://developer.twitter.com/en/docs/tweets/filter-realtime/overview.html", 2021. [Accessed: 2021-05-07].

[239] Twitter. Q3 2021 - Letter to Shareholders, 2121.

[240] Aman Tyagi and M Kathleen. Climate Change Conspiracy Theories on Social Media. *preprint*, pages 2–11, 2021.

[241] R Valecha, T Volety, H R Rao, and K H Kwon. Misinformation Sharing on Twitter During Zika: An Investigation of the Effect of Threat and Distance. *IEEE Internet Computing*, 25(1):31–39, 1 2021.

[242] Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, page 146144481771208, 2017.

[243] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 280–289, 2017.

[244] Nikhita Vedula and Srinivasan Parthasarathy. FACE-KEG: Fact Checking Explained using KnowledgE Graphs. *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 526–534, 2021.

[245] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. *55th Annual Meeting of the Association for Computational Linguis*, pages 647–653, 7 2017.

[246] Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data*, 11(4), 2017.

[247] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[248] Mason Walker and Katerina Eva Matsa. News consumption across social media in 2021. *Pew Research*, 20:2021, 2021.

[249] Benjamin X Wang and Nathalie Japkowicz. Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 25(1):1–20, 2010.

[250] Hui Wang, Lin Deng, Fei Xie, Hui Xu, and Jianghong Han. A new rumor propagation model on SNS structure. *Proceedings - 2012 IEEE International Conference on Granular Computing, GrC 2012*, pages 499–503, 2012.

[251] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426. Association for Computational Linguistics, 2017.

[252] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857, 2018.

[253] C. Nadine Wathen and Jacquelyn Burkell. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 53(2):134–144, 2002.

[254] Ke Wu, Song Yang, and Kenny Q. Zhu. False rumors detection on Sina Weibo by propagation structures. *Proceedings - International Conference on Data Engineering*, 2015-May:651–662, 2015.

[255] Liang Wu, Fred Morstatter, Xia Hu, and Huan Liu. Mining Misinformation in Social Media. In *Big Data in Complex and Social Networks*, chapter 5. Taylor & Francis, New York, 2016.

[256] Cao Xiao, David Mandell Freeman, and Theodore Hwa. Detecting Clusters of Fake Accounts in Online Social Networks. *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security - AISec '15*, pages 91–101, 2015.

[257] Bailin Xie, Yu Wang, Chao Chen, and Yang Xiang. Gatekeeping behavior analysis for information credibility assessment on weibo. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[258] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. TURank: Twitter user ranking based on user-tweet graph analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6488 LNCS:240–253, 2010.

[259] Fan Yang, Xiaohui Yu, Yang Liu, and Min Yang. Automatic detection of rumor on Sina Weibo. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2, 2012.

[260] Kai Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. *arXiv*, 2020.

[261] Zhenhuang Yong, Hanbing Yao, and Yefu Wu. Rumors Detection in Sina Weibo Based on Text and User Characteristics. In *Proceedings of 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2018*, 2018.

[262] Razieh Nokhbeh Zaeem, Chengjing Li, and K. Suzanne Barber. On Sentiment of Online Fake News. *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020*, pages 760–767, 2020.

[263] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, January 2019.

[264] Qiao Zhang, Shuiyuan Zhang, Jian Dong, Jinhua Xiong, Xueqi Cheng, Juanzi Li, Heng Ji, Dongyan Zhao, and Yansong Feng. Automatic Detection of Rumor on Social Network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9362:113–122, 2015.

[265] Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, 2021.

[266] Xinyi Zhou and Reza Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5), 2020.

[267] Fabiana Zollo and Walter Quattrociocchi. Misinformation Spreading on Facebook. In Sune Lehmann and Yong-Yeol Ahn, editors, *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, pages 177–196. Springer International Publishing, Cham, 2018.

[268] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, 2015.

[269] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 2016.