TOWARDS AUTOMATIC FAKE NEWS DETECTION IN DIGITAL PLATFORMS: PROPERTIES, LIMITATIONS, AND APPLICATIONS

JULIO CESAR SOARES DOS REIS

TOWARDS AUTOMATIC FAKE NEWS DETECTION IN DIGITAL PLATFORMS: PROPERTIES, LIMITATIONS, AND APPLICATIONS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Fabrício Benevenuto de Souza

Belo Horizonte
Novembro de 2020

JULIO CESAR SOARES DOS REIS

TOWARDS AUTOMATIC FAKE NEWS DETECTION IN DIGITAL PLATFORMS: PROPERTIES, LIMITATIONS, AND APPLICATIONS

Thesis presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte
November 2020

Reis, Julio Cesar Soares dos.

R375t Towards automatic fake news detection in digital platforms [manuscrito]: properties, limitations, and applications / Julio Cesar Soares dos Reis. – 2020.
xxx, 194 f. il.

Orientador: Fabrício Benevenuto de Souza.
Tese (Doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.103-132

1. Computação – Teses. 2. Mídia social - Teses. 3. Fake news - Teses. 4. Desinformação - Teses. 5. Aprendizagem de máquina - Teses I. Souza, Fabrício Benevenuto de. II. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas, Departamento de Ciência da Computação. III.Título.

CDU 519.6*04(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Towards Automatic Fake News Detection in Digital Platforms: Properties, Limitations, and Applications

JULIO CESAR SOARES DOS REIS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Fabrício Benevenuto

Prof. Fabrício Benevenuto de Souza - Orientador
Departamento de Ciência da Computação - UFMG

Profa. Viviane Pereira Moreira
Departamento de Informática Aplicada - UFRGS

Prof. Leandro Balby Marinho
Departamento de Sistemas e Computação - UFCG

Prof. Fabrício Murai Ferreira
Departamento de Ciência da Computação - UFMG

Profa. Mirella Moura Moro
Departamento de Ciência da Computação - UFMG

Prof. Adriano Alonso Veloso
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 3 de Novembro de 2020.

To my beloved Grandfather, José Francisco dos Reis (in Memoriam).

Acknowledgments

Primeiramente eu gostaria de agradecer a Deus. Foi Ele quem me permitiu a realização deste trabalho, estando comigo em todos os instantes.

Depois, eu gostaria de agradecer a minha esposa Lidiana e aos meus filhos Miguel e Maria Alice, pela paciência e compreensão nos vários momentos em que eu não pude estar presente.

Aos meus pais, Meire e Robson, aos meus irmãos Paulo e Caio, e demais familiares e irmãos da IEQ Itatiaia que me deram todo o suporte necessário para que essa jornada fosse possível, sempre me incentivando na busca pelo sucesso pessoal e acadêmico.

Especialmente, eu gostaria de agradecer ao meu orientador, o professor Fabrício Benevenuto, pelo apoio, paciência, e conhecimento compartilhado, essenciais durante este projeto. Seu entusiasmo é contagiante! Estendo também este agradecimento a todos os professores e coautores dos trabalhos dos quais tenho participado durante os últimos anos.

Aos meus amigos e colegas do LOCUS, PENSI e em última instância do DCC, pela rica troca de experiências e, especialmente, aqueles que estiveram comigo durante o período de qualificação do doutorado. Com eles pude compartilhar momentos de dificuldade e felicidade, o que tornou essa trajetória possível.

Também gostaria de agradecer ao apoio financeiro oferecido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e pelo Google por meio do projeto LARA (*Latin American Research Awards* - Edição 2018). Por fim, eu gostaria de agradecer, especialmente à todos que de alguma forma estão contribuindo para a construção desta pesquisa, de forma direta ou não, com participação nas diversas etapas deste projeto.

Muito obrigado a todos! Que o Eterno os abençoe imensamente!

"A tarefa não é tanto ver aquilo que ninguém viu,
mas pensar o que ninguém ainda pensou
sobre aquilo que todo mundo vê."
(Arthur Schopenhauer)

Resumo

As plataformas digitais mudaram drasticamente a forma como as notícias são produzidas, disseminadas e consumidas em nossa sociedade. Um problema fundamental hoje é que as plataformas digitais se tornaram espaços amplamente abusados por campanhas de desinformação que afetam a credibilidade de todo o ecossistema de notícias. O surgimento de notícias falsas nesses ambientes evoluiu rapidamente para um fenômeno mundial, onde a falta de estratégias escaláveis de verificação de fatos é preocupante. Assim, soluções automáticas para detecção de notícias falsas poderiam ser usadas por jornalistas e equipes de checagem de fatos como uma ferramenta auxiliar na identificação de notícias com alta probabilidade de serem falsas. Neste contexto, esta tese tem como objetivo investigar abordagens práticas para a detecção automática de notícias falsas disseminadas em plataformas digitais. Para isso, inicialmente nós pesquisamos um grande número de trabalhos recentes e relacionados como uma tentativa de implementar atributos propostos na literatura para a detecção de notícias falsas. Isso nos possibilitou propor novos recursos, explorar conjuntos de dados rotulados disponíveis e propor um novo conjunto de dados para avaliar o desempenho de previsão das atuais abordagens de aprendizado de máquina supervisionadas na realização desta tarefa. Nossos resultados revelam que esses modelos computacionais propostos possuem um grau útil de poder discriminativo para detectar notícias falsas disseminadas em plataformas digitais. Além disso, nós propomos um arcabouço imparcial para quantificar a informatividade de atributos para detecção de notícias falsas. Como parte de nosso arcabouço proposto, apresentamos uma explicação dos fatores que contribuem para as decisões do modelo, promovendo assim o raciocínio cívico, complementando nossa capacidade de avaliar o conteúdo digital e chegar a conclusões justificadas. Também analisamos recursos e modelos que podem ser úteis para detectar notícias falsas em diferentes cenários: eleições nos Estados Unidos e no Brasil. Por fim, propomos e implementamos em um sistema real um novo mecanismo que, conforme resultados experimentais, reduziu significativamente o número de notícias que jornalistas e verificadores de fatos precisam ler antes de encontrar uma história falsa.

Palavras-chave: Plataformas Digitais; Mídia Social; Notícias falsas; Detecção de Notícias Falsas; Desinformação; Checagem de Fatos; Atributos; Aprendizagem de Máquina; Interpretabilidade de Modelos; Informatividade de Atributos.

Abstract

Digital platforms have dramatically changed the way news is produced, disseminated, and consumed in our society. A key problem today is that digital platforms have become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem. The emergence of fake news in these environments has quickly evolved into a worldwide phenomenon, where the lack of scalable fact-checking strategies is especially worrisome. Thus, automatic solutions for fake news detection could be used as an auxiliary tool for fact-checkers to identify content that is more likely to be fake, or content that is worth checking. In this context, this thesis aims at investigating practical approaches for the automatic detection of fake news in digital platforms. First, we survey a large number of recent and related works as an effort to implement all potential features to detect fake news. We propose novel features and explore labeled datasets proposing new ones to assess the prediction performance of current supervised machine learning approaches. Our results reveal that these proposed computational models have a useful discriminative capacity for detecting fake news disseminated in digital platforms. We then propose an unbiased framework for quantifying the informativeness of features for fake news detection. As part of our proposed framework, we present an explanation of factors contributing to model decisions, thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions. We also analyze features and models that can be useful for detecting fake news from different scenarios: the US and Brazilian elections. Finally, we propose and implement into a real system a new mechanism that accounts for the potential occurrence of fake news within data, significantly reducing the number of content pieces journalists and fact-checkers have to go through before finding a fake story.

Palavras-chave: Digital Platforms; Social Media; Fake News; Fake News Detection; Misinformation; Fact-Checking; Features; Machine Learning; Explainable Models; Informativeness of Features.

List of Figures

List of Tables

List of Abbreviations and Acronyms

Acc. *Accuracy*

API *Application Programming Interface*

AUC *Area Under the Curve*

ASN *Autonomous System Number*

FPR *False Positive Rate*

IG *Information Gain - InfoGain*

IP *Internet Protocol*

IRA *Russian Intelligence Research Agency*

KNN *k-Nearest Neighbors*

LIWC *Linguistic Inquiry and Word Count*

MAD *Mean Absolute Deviation*

NB *Naive Bayes*

NDCG *Normalized Discounted Cumulative Gain*

OCR *Optical Character Recognition*

P *Precision*

PII *Personally Identifiable Information*

PoS tagging *Part-of-Speech Tagging*

R *Recall*

RG *Research Goal*

RF *Random Forests*

ROC Curve *Receiver Operating Characteristics Curve*

SVM *non-linear Support Vector Machine with the Radial Basis Function*

TPR *True Positive Rate*

t-SNE *t-Distributed Stochastic Neighbor Embedding*

URL *Uniform Resource Locator*

US *United States*

USA *United States of America*

XGB *eXtreme Gradient Boosting - XGBoost*

Contents

Chapter 1

Introduction

1.1 Motivation

Digital platforms, including social media systems and messaging applications, are actively used by over one-third of the world's population [91]. These platforms have significantly changed the way users interact and communicate online, opening a whole new wave of applications, and modifying existing information ecosystems. Particularly, digital platforms have dramatically changed the way news is produced, disseminated, and consumed, opening unforeseen opportunities, and also creating complex challenges.

Part of the reasons for this change are inherent to the nature of these digital platforms: (i) it is often more timely and less expensive to produce and consume news on digital platforms compared with traditional news media, such as newspapers or television; and (ii) it is easier to share, comment on, and discuss the news with friends or other readers in digital platforms, which enhances communication and interactions among users [246]. Hence, digital platforms are shaping the way users consume information. Nowadays, about 62% of US users and 66% of Brazilian users get news from digital platforms [170, 214]. Despite the numerous benefits that these systems bring to our society, they have become a place for campaigns of misinformation which are often intended to deceive people, especially in contexts such as health and politics.

Regarding health, the flood of fake medical news disseminated on digital platforms is causing irreparable damage [63]. For instance, a cancer patient mistook an online ad for experimental cancer treatment as medically reliable information, which resulted in his death [63][1]. Furthermore, during the COVID-19 pandemic, there has been an uptick in rumours and conspiracies spreading through social platforms [82]. The International

[1] https://www.bbc.com/news/business-36189252

Fact-Checking Network found more than 3,500 false claims related to COVID-19 in less than two months [203]. As result, at least 800 people may have died around the world because of coronavirus-related misinformation in the first three months of 2020[2].

In the political context, election after election, we can see different forms of misconduct and complex strategies of opinion manipulation through the spread of fake news. The 2016 presidential election in the USA is still remembered for a 'misinformation war' that happened mostly through Twitter and Facebook. The notorious case involved an attempt of influence from Russia through targeting advertising [221]. Similar attempts were observed during the 2018 Brazilian elections, where WhatsApp was abused to send out misinformation campaigns, with large use of manipulated images and memes[3] containing all kinds of political attacks. A recent study showed that 88% of the most popular images shared in the last month before the Brazilian elections were fake or misleading [258]. Also using WhatsApp, in India, fake rumors spread through the online service were responsible for multiple cases of lynching and social unrest [17].

A unique characteristic of news in digital platforms that supports this phenomenon of fake news is that anyone can register/behave as a news publisher without any upfront cost (e.g., anyone can create a Facebook page claiming to be a newspaper or news media organization, or yet, create a group on WhatsApp to spread news). Consequently, not only traditional news corporations are increasingly migrating to digital platforms, but also many news outlets are also emerging on these environments[4]. For instance, previous efforts showed that in 2018 there were more than 20 thousand pages in the USA categorized as news publishers on Facebook [219], and this number is continuously growing.

Along with this transition, there are growing concerns about fake news publishers producing and posting fake news stories[5], and often disseminating them widely through social platforms [139]. For instance, a study funded by Avaaz[6] asked Brazilian voters whether they saw and believed in five of the most popular fake news on digital platforms during the last weeks of the election in 2018. Impressively, the results revealed that over 98% of interviewed voters were exposed to one or more fake news articles and

[2] https://www.bbc.com/news/world-53755067

[3] "An image, video, piece of text, etc., typically humorous in nature, that is copied and spread rapidly by Internet users, often with slight variations" [192].

[4] https://www.comscore.com/Insights/Blog/Traditional-News-Publishers-Take-Non-Traditional-Path-to-Digital-Growth

[5] https://www1.folha.uol.com.br/ilustrissima/2017/02/1859808-como-funciona-a-engrenagem-das-noticias-falsas-no-brasil.shtml

[6] Avaaz (www.avaaz.org/) is the campaigning community bringing people-powered politics to decision making worldwide.

almost 90% of them believed that these stories were true[7]. Potentially, these numbers impacted the democracy in Brazil face of the 2018 presidential elections.

Misinformation, spin, lies, and deceit have been around forever, but the rise of digital platforms has potentially increased the spread of misinformation and thus have turned the problem of fake news into a worldwide phenomenon, where the lack of scalable fact-checking strategies is especially worrisome. Therefore, in this scenario, our hypothesis is that automatic fake news detection can have a useful degree of discriminative power to identify content that is more likely to be fake, supporting the fact-checking process as well as minimizing the impact caused by extensive production, dissemination, and consumption of fake news through digital platforms.

1.2 Thesis Statement

In this thesis, we aim at investigating the potential of automatic solutions to identify fake news disseminated on digital platforms. Whereas fact-checking is an essential strategy to identify fake news that is simple but does not scale, automatic solutions for fake news detection could be used as an assistive tool for fact-checkers to identify content that is more likely to be fake or content that is worth checking, still leaving the final call to an expert at the endpoint of the process. Furthermore, these strategies could be incorporated by digital platforms and search engines as a way to limit the audience of suspicious news stories.

However, automatically identifying fake news is not a trivial task. First, humans themselves are naturally limited at differentiating between real and fake news [246], especially when it comes to sensitive subjects, such as politics and health. In addition, news stories are produced by different sources in which each one has its own content style and intrinsic bias, and they are disseminated in different ways through distinct environments, which makes the fake news identification task even harder. Thus, each of these aspects of news (i.e., content, source, environment) can be modeled according to a different set of features that can allow an understanding of typical patterns of fake news that hold across different scenarios. Assessing those differences is crucial to enable the development of language/culture agnostic models for fake news detection.

Therefore, we intend to explore features and solutions that remain useful considering different scenarios and investigate strategies with practical potential for detecting fake news spread on digital platforms.

[7] https://www1.folha.uol.com.br/poder/2018/11/90-dos-eleitores-de-bolsonaro-acreditaram-em-fake-news-diz-estudo.shtml

1.3 Research Goals

The general objective of this thesis can be divided into the following specific research goals (RGs).

- **RG1 - Assessing the Prediction Performance of Solutions to Detect Fake News:** There are some current research efforts aiming to understand fake news phenomena and to identify typical patterns and features for proposing automatic solutions for fake news detection [53, 257, 276, 282]. Despite the undeniable importance of the existing efforts in this direction, they are mostly concurrent work which identifies recurrent patterns on fake news after they have been already disseminated or that propose new features for training classifiers using data from a specific scenario, based on ideas that have not been tested in combination. Thus, it is difficult to gauge the practical potential that supervised models trained from features proposed in recent studies have for detecting fake news. Hence, we use available data and build a new dataset for exploiting the main features proposed in the literature for fake news detection and propose new ones, to evaluate and compare different supervised machine learning approaches, assessing their prediction performance in the task of automatically identify fake news disseminated in different scenarios. Particularly, we use data from two political events that have been extensively abused by misinformation campaigns [26, 30, 216]: (i) the 2016 US presidential election, and (ii) the 2018 Brazilian presidential election. Furthermore, we also explore data from the health context in order to compare our results and measure the potential of our features for fake news detection in a scenario different from politics. Specifically, we run experiments to answer the following question: *What is the prediction performance of current approaches and features for automatic detection of fake news considering data from different scenarios?*

- **RG2 - Quantifying the Informativeness of Features for Fake News Detection:** Another open issue is that little is known about the discriminating power of features proposed in the literature for fake news detection, either individually or when combined with others, especially involving different scenarios. Some may be adequate for pinpointing fake news with specific patterns, while others are more general but not sufficiently discriminating. Moreover, while explaining the decisions made by the proposed algorithms for fake news detection is central to understand the structure of fake content, this discussion is often left aside. We address all these issues in this research step. Specifically, after assessing the prediction performance of current supervised machine learning approaches and features for automatic detection of fake news, we provide answers to the following questions. *Do we need all proposed features for fake news detection, or should we focus on a smaller set of more representative features? Is there a trade-off between feature discriminating power and robustness to pattern variations? Is there a clear link between features and the patterns of fake news they can detect?* Since the considered features for fake news detection may have a variety of complex nonlinear interactions, we propose a framework for quantifying their informativeness. In addition, we build models employing a fast and effective classification algorithm with significant flexibility and propose an unbiased strategy to generating them, which enables to perform a unique macro-to-micro investigation of the considered features. We hypothesize that there is no single model to tackle all facets of fake news detection, suggesting that understanding the informativeness of specific combinations of features can be useful for building robust models capable of identifying fake news with different patterns. To accomplish this task, we also explore the data from different scenarios as introduced in the first research goal. As part of our proposed framework, we also present an explanation of factors contributing to model decisions, thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions. Last, we investigate whether *there is a set of features that yield models with high performance and able to identify fake news disseminated on digital platforms considering data from different scenarios*, i.e., the 2016 US and 2018 Brazilian presidential elections. In order to accomplish such a goal, we propose an experiment based on Pareto-Efficiency [295], which is a central concept in Economics widely explored in several areas of knowledge, including the Computer Science [150].

- **RG3 - Exploring the Practical Potential of Fake News Detection**: We explore our findings towards automatic fake news detection to develop a new strategy to help fact-checkers identify news stories that are more likely to be fake, incorporating our approach into a real system called the WhatsApp Monitor (`http://www.whatsapp-monitor.dcc.ufmg.br/`). Particularly, this is a web-based system proposed by our research group that helps researchers and journalists by ranking content shared on WhatsApp public groups and displaying them in an organized way. Even a simpler version of the system that daily updates the most popular content shared on WhatsApp proved to be very useful as, during the Brazilian 2018 elections, over a hundred journalists and three fact-checking agencies had access to it and explicitly mentioned our system as a data source [167]. Although this system has already been extensively used, it only displays a list of the most frequently shared content in the monitored groups over a time interval. This does not necessarily indicate which content should be fact-checked first, as the popularity of a news story in WhatsApp may not be representative of its popularity elsewhere. Therefore, we propose and implement a new mechanism that accounts for the potential occurrence of fake news within our data, significantly reducing the number of content pieces journalists and fact-checkers have to go through before finding a fake story. Specifically, we use the supervised machine learning methods explored in this thesis to estimate a *fakeness score* on news stories aiming at improving ranking results, which can support decisions regarding the selection of facts (or news) to be checked. Last, we deploy our approach in the WhatsApp Monitor.

1.4 Contributions

The main contributions of this thesis are summarized as follows.

1. A survey that describes datasets and features for fake news detection (RG1). We conduct a systematic survey that includes identifying existing features and datasets for the fake news detection task as an effort to implement and compare them;

2. A new dataset for fake news detection (RG1). We perform an extensive data collection from a large set of WhatsApp public groups and the websites of fact-checking agencies in Brazil to build a new dataset containing fact-checked fake images shared through WhatsApp during the Brazilian elections in 2018;

3. A new set of useful features for fake news detection (RG1). In addition to exploring the main features proposed in the literature for fake news detection, we propose novel features, such as those related to the properties of images associated with the news story, the semantic structure of news text, news sources (e.g., location and credibility of publishers), and new propagation measures within/outside digital platforms, which have a promising discriminative degree to distinguish fake news stories from others;

4. A measurement of the prediction performance of current approaches and features for automatic detection of fake news disseminated on digital platforms (RG1). We

explore datasets from different scenarios and features for fake news detection to assess the ability of supervised machine learning approaches to correctly identify fake news stories;

5. A framework for quantifying the informativeness of features for fake news detection (RG2). We propose an unbiased strategy for models generation where each model is composed of a set of randomly chosen features, enabling us to perform a deep investigation of the informativeness of features for detecting fake news. The investigation unveils the real impact of a sleigh of features for fake news detection highlighting how hard is fake news detection and evidence that different combinations of features are tailored for detecting fake news with different patterns;

6. An explanation of factors contributing to fake news model decisions (RG2). As part of our proposed framework, we use a state-of-the-art technique to explain why news are classified as fake or not by fake news detection models thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions;

7. An investigation of the features and models that can be useful for detecting fake news in different scenarios (RG2). We perform an analysis to contrast the informativeness of features for fake news detection considering data from different scenarios, i.e., the 2016 US and 2018 Brazilian presidential elections;

8. A new strategy with the practical potential for detecting fake news spread on digital platforms deployed in a real system (RG3). We design and integrate a new approach to the WhatsApp Monitor system that allows users to rank news stories disseminated through images according to an estimated *fakeness score*, helping the fact-checking task.

1.5 Chapters Organization

The rest of this thesis is organized as follows. Chapter 2 states the background for developing this thesis as well as its related work. Then, Chapter 3 presents the datasets used for the fake news detection task including our strategy to build a new one. We describe existing features for fake news detection and how we implemented them, including our new proposed ones in Chapter 4. In turn, the datasets and features for fake news detection that are the base of our measurements of prediction performance of current machine learning approaches are described in Chapter 5. Chapter 6 presents our framework proposed to quantify the informativeness of features for fake news detection including our results regarding features and models that can be useful for detecting fake news in different scenarios. Then, we describe our practical strategy to help fact-checkers identify fake news disseminated on digital platforms in Chapter 7. Finally, Chapter 8 concludes this thesis and discusses future work.

Chapter 2

Background and Related Work

In this chapter, we present a summary of background information and related work that is fundamental to the understanding of this thesis. We discuss previous efforts related to each of the aforementioned research goals (RGs), which can be grouped into several topics, as shown in Figure 2.1.

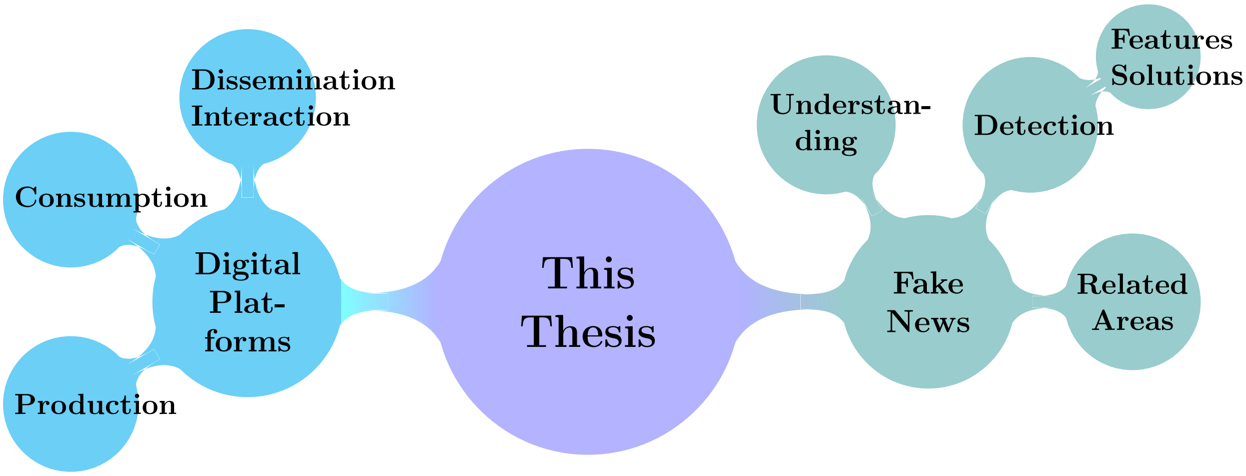


Figure 2.1: Mind map of the themes related to this thesis.

2.1 News Media Ecosystem in Digital Platforms

News media have long been a subject of studies by scholars in various domains, such as journalism, communication, and political science. Nonetheless, since news media have tapped into Web and digital platforms, it has also been a topic of interest by computer scientists. With the emergence of the digital era, news media have started publishing in the digital medium. Hence, with an abundant digital trace of news information, the possibilities of new applications, and the emergence of new challenges in this complex scenario, computer scientists have investigated problems related to the news ecosystem in digital platforms, but usually with different goals and aims.

Broadly speaking, the basis of the news ecosystem in digital platforms can be divided into three main components: **production**, **consumption**, and **dissemination and interaction**, as shown in Figure 2.2. First, before digital platforms, news articles were produced (or written) only by traditional news media organizations (i.e., newspapers) or independent journalists. With the rise of digital platforms, a key characteristic of news **production** in these environments is that anyone can be a news producer (e.g., anyone can create a user on a digital platform to produce and spread news without any upfront cost). Additionally, the **consumption** of news also changed over time from newsprint to radio/television and, recently, online news and digital platforms, where it is often more timely and less expensive to consume news compared with traditional news media. For instance, a survey by Pew Research Center estimates that 62% of the adults in the US consume news primarily from social media sites [170]. In Brazil, according to a survey conducted by the Reuters Institute, this percentage reaches 66% [214]. Last, digital platforms introduce new mechanisms for information **dissemination and interaction**, allowing users to share and promote news stories according to their will.

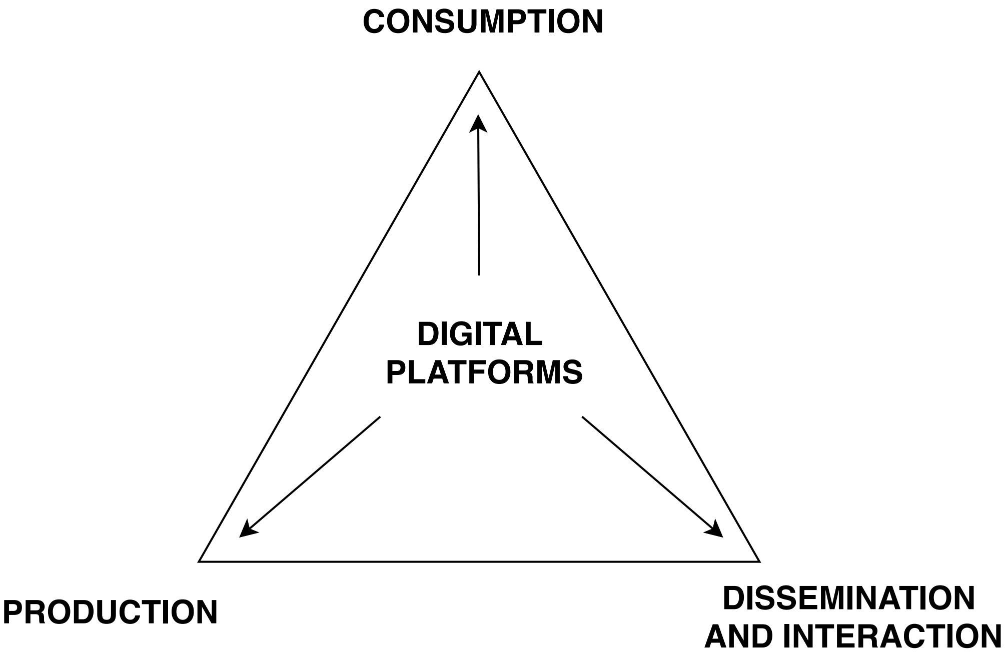


Figure 2.2: News ecosystem in digital platforms.

As a result of the insertion of digital platforms in the news ecosystem, there are different Computer Science efforts that attempt to better understand these changes and propose solutions to support this phenomenon in its various stages. These efforts can be grouped into three related sets to the main components of the news ecosystem in digital platforms. The first covers content **production** and involves topics such as news coverage [132], news events [2, 149, 210, 293], credibility [120], aspects of news attractiveness [126], news bias [33, 58], etc. The second is related to content **consumption** and involves users reading patterns [55, 129], strategies to provide users better information diets [124], personalization [65, 93], summarization [92], recommendation systems [178], content visualization [244], and news consumption on mobile devices [54, 287]. The third is associated with **dissemination and interaction** mechanisms provided by digital platforms, including efforts on understanding these mechanisms [41], motivations for users to share [140] and new challenges that emerge from these mechanisms such as filter bubbles and echo chambers [22]. Finally, these sets can be subdivided into smaller groupings of studies, which we present next. An overview of such subsets is presented in Figure 2.3.

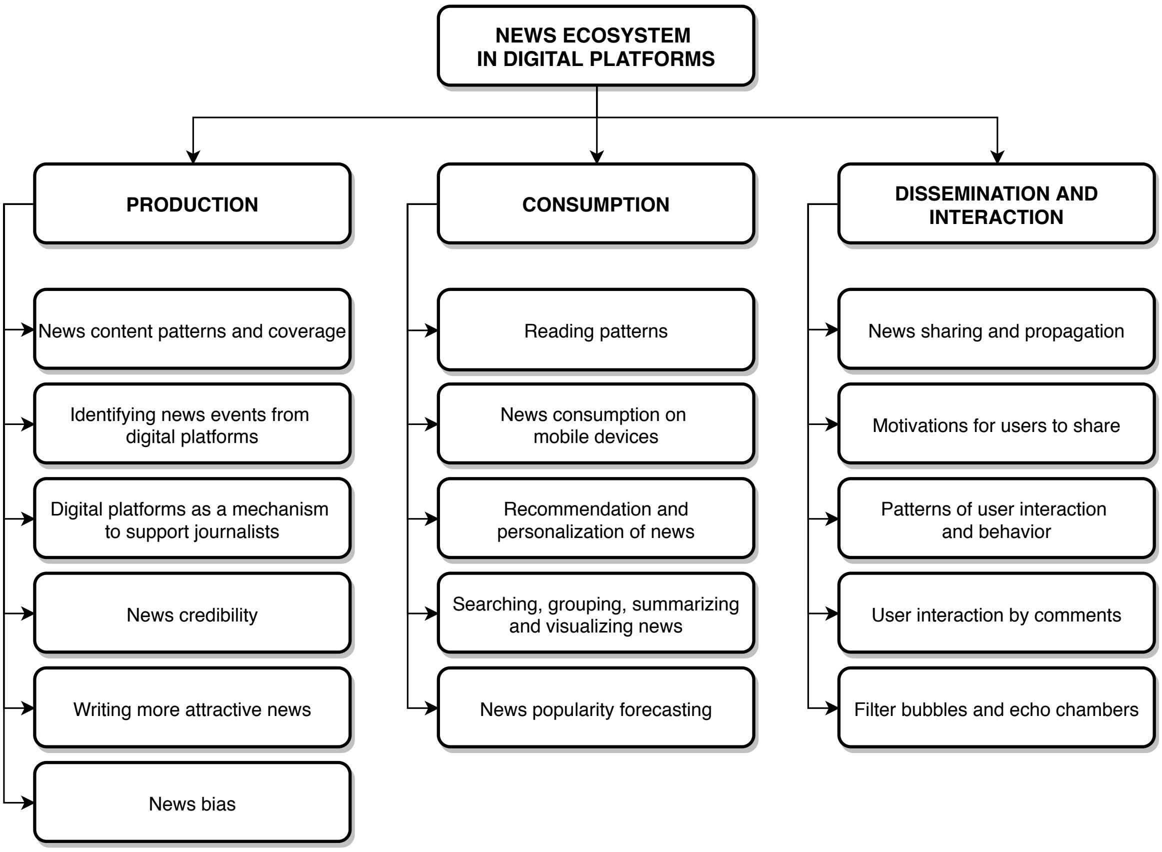

NEWS ECOSYSTEM IN DIGITAL PLATFORMS
PRODUCTION
CONSUMPTION
DISSEMINATION AND INTERACTION
News content patterns and coverage
Identifying news events from digital platforms
Digital platforms as a mechanism to support journalists
News credibility
Writing more attractive news
News bias
Reading patterns
News consumption on mobile devices
Recommendation and personalization of news
Searching, grouping, summarizing and visualizing news
News popularity forecasting
News sharing and propagation
Motivations for users to share
Patterns of user interaction and behavior
User interaction by comments
Filter bubbles and echo chambers

Figure 2.3: Overview of efforts related to the news ecosystem in digital platforms.

2.1.1 Production

We start by reviewing previous efforts that explore the dynamics of online news production, and how this process has been affected by the inclusion of digital platforms in the news ecosystem.

News content patterns and coverage. News content patterns are explored by various studies [213], focusing on different elements, such as photos [133], images [196] and news videos [200]. There are also approaches to estimate the impact of news content on digital platforms. Such approaches merge natural language processing techniques and are capable of providing indicators of the social platform's impact of the specified content, comparing their relevance to an arbitrary news article [68]. In addition, some analyses show that certain characteristics of the news content can impact the dynamics of the digital platforms, for example, how faster bad news travel than good news on Twitter [180]. Thus, large scale datasets of online news are also used to understand how features extracted from news articles are related to local [195] and global news coverage, focusing on different aspects, such as disasters [132], and epidemics and pandemics [109]. Yet, the news coverage in digital platforms is often dependent on how they are framed-typically by mainstream media [188]. Hence, there is still a gap between what is published by online news media, what online news media publish, and what the general public produces on digital platforms.

Identifying news events from digital platforms. Digital platforms services are being used extensively as news sources and for spreading information on real-world events. Thus, a large volume of news-related research by computer scientists focus on the automated identification of events based on digital platforms activity [2, 149, 210]. These efforts range from sifting through all information noise on digital platforms for detecting and verifying the veracity of news events and emerging topics [4, 115, 153] to proposing real-time systems that make an editorial decision about the level of accuracy and interest for a given news event [238]. Overall, all these technological developments coupled with the proliferation of information sources through digital platforms created new opportunities that are revolutionizing the practice of journalism in our society [238].

Digital platforms as a mechanism to support journalists. Digital platforms have changed many aspects of news production, including the relationship of journalists with their work [141, 175]. The fourth annual Digital Journalism Study of the 2011[1] interviewed approximately 500 journalists from 15 different countries and found that almost 50% of them used social media as a primary information medium for writing their news. Increasingly, they are using social media as resources for the identification of stories [72, 142], to improve (or enrich) their articles [71] and last, to disseminate news they wrote [190]. Not surprisingly, recent research efforts are devoted to proposing technological mechanisms that can support and help these journalists in these environments [265].

News credibility. Journalists, fact-checkers, and common readers are facing the challenge of discovering credible news from more diverse and unreliable information in the age of digital platforms. More and more news events break on digital platforms first and are picked up by news media subsequently. In recent years, with the huge popularity of systems like Twitter, Facebook, and WhatsApp, digital platforms have been a powerful channel to spread misinformation. There are two kinds of efforts to tackle this problem. First, some studies aim at better comprehending the phenomenon [139, 207, 277]. Studies show that misinformation tends to spread faster than the real news on digital platforms [277], evidencing the complexity of the problem in this environment. Second, there are studies that propose mechanisms to evaluate news veracity [120, 276], including efforts that explore fact-checking as an effective way to contain misinformation spreading on social platforms [50, 110, 111]. This topic is further discussed in the next sections.

Writing more attractive news. Online readers are often willing to spend a limited amount of time for consuming news, therefore, it is critical for news sites to have effective strategies to catch people's attention and attract their clicks, through the use of attractive words [126], visual elements (layout, color, photographs, and front page) [77], and/or effective clickbaits [44]. Previous efforts defined a simple method to measure the click-value of individual words and analyze how temporal trends and linguistic attributes affect the click-through rate of news articles [126]. As a result, some studies show that identifying or discovering adequate words for headlines can be useful to generate more clicks in news articles in the future [126].

News bias. Readers of news on digital platforms are often not aware of the biases of the newspapers [219]. For traditional media, two broad strategies have been used to quantify the biases of a given news outlet: (i) the first class of approaches quantifies media bias directly by inspecting the published content [33, 58, 233], specifically focusing on the coverage of important events by the media organizations, and (ii)

[1] http://www.oriellaprnetwork.com/

the second strategy is to analyze the readership of the news outlets, which assumes that the content and attitudes of a news outlet end up driving the biases of its audience [22, 97, 304]. Based on that, recent efforts explore scalable strategies to accurately and automatically infer the biases of thousands of news sources on digital platforms such as Facebook and Twitter [219, 255], and analyze how these bias can be visualized and communicated [254].

2.1.2 Consumption

The emergence of digital platforms as a new way to produce, disseminate, and consume news, enables users to find and consume information in an unmediated way [69]. Thus, some emerging studies aim to explore aspects of news consumption in social media, including reading patterns, mechanisms to support consumption, besides strategies to encourage reading, such as recommendation and visualization of news.

Reading patterns. Reading patterns in online news portals were widely exploited [137], including the analysis of gender and age differences [9], and the role of geographic information in news consumption [96]. Nonetheless, the news consumption from digital platforms is different [112]. There are many news articles distributed, and now, the news providers are under pressure to find ways of engaging the attention of readers [126, 129], impacting, inclusive, on the news production. Thus, there are studies that attempt to characterize reading habits of online news readers focusing on aspects related to freedom of reading choices offered by digital platforms [11] and explore how reading patterns can be used to provide insights for the better design of news recommendation systems [79, 272] and yet, to infer the quality of the text [253].

News consumption on mobile devices. Mobile devices revolutionized information access, facilitating their consumption since it is possible to do it from anywhere and at a low cost. Particularly, such a change has affected the news ecosystem [270], especially regarding their consumption. Accessing mobile news has gained traction in the everyday life of the readers [287]. Thereby, some studies are emerging to understand aspects of the consumption of online news through mobile devices [181], and to propose new apps to support them [54]. These apps are capable of unobtrusive logging of news interactions and recognizing patterns of user's news reading behavior [55], personalizing [93] and improving their experience in this scenario.

Recommendation and personalization of news. Recommending news to readers can be an effective strategy to encourage their consumption. Thus, there are some approaches focused on identifying what users usually are interested in reading [1], for instance, based on their previous behavior sequence [75] or past clicks [151]. On the other hand, few strategies explore diversity on recommendation [236] as a way to avoid filter bubbles [8, 22]. Last, studies show that personalizing news digest can improve the user experience during news reading, potentializing their engagement in this task [42, 65, 124, 145].

Searching, grouping, summarizing and visualizing news. There is a lot of news being generated every day. Thus, searching for news of interest can be a costly task for users. From this, some strategies have emerged to facilitate the search for news [178], focusing on a certain subject [279], content feature (e.g., polarity) [76], or yet, using user-generated content themselves [165]. Also, there are some emerging approaches that attempt to summarize trending subjects by jointly discovering the representative and complementary information from news and posts [92]. Furthermore, it is important to help users quickly understand and act upon the large volume of data [244] and another effective way to do so is to visualize it. Thus, some studies aim to propose strategies for displaying news stories in a hierarchical map [189, 262] that is generated automatically working as a tool for browsing online news [189]. In addition, users can navigate through based on the news topics or on the geo-locations, for instance. In this way, some efforts show a real-time Twitter-based application that visualizes an ideological map of various media sources [8].

News popularity forecasting. There has been an increase in scientific interest in discovering news articles that may become popular among users [179, 259, 280]. Some studies investigate the relationship of this popularity with other aspects of the news, such as the sentiment of the headlines [213]. The task of predicting the popularity of news can be explored through the application of ranking techniques [259], or applying regression and classification algorithms [23], focusing in temporal aspects [144], social dynamics [143], or yet comments as a measure of popularity [260]. In digital platforms, factors as context, network properties, and content have made the task of predicting the popularity of news articles to become even more challenging task [23].

2.1.3 Dissemination and Interaction

As digital platforms become an important channel for news diffusion, some efforts attempted to investigate how news are shared in these systems. In this context, we also conduct an exploratory analysis focusing on the understanding of a key component of this process: the news spreaders, which are people who share news stories through digital platforms posts. We show that they play an important role in expanding the audience of news on Twitter, which would otherwise be very limited. These results are presented in the Appendix A. Next, we detail some other studies that provide interesting findings in this field, highlighting some of them that exploit user engagement (e.g., by comments) along with the news dissemination process, including their motivations and behavior dynamics. Last, we discuss filter bubbles and the echo chambers' effect in the news ecosystem.

News sharing and propagation. The dissemination of news in digital platforms has been the subject of several studies [28] focusing on different aspects, such as bias [41] and political news [11], or yet, the characteristics of players (or spreaders) [116] and their role in such propagation process [39, 40, 225, 289]. Looking at Twitter, for example, some studies show that retweets are responsible for increasing the audience of URLs by about two orders of magnitude [224]. Moreover, there are several research efforts that attempt to understand characteristics (or factors) influencing news sharing in digital platforms [140]. Some of them, for instance, show that bad news tends to spread faster in systems such as Twitter [180]. Also, a recent effort [32] has tackled the question "Why are some news articles shared more than others?". The author shows that story importance cues are relevant in driving social sharing and that certain topics (i.e., stories about politics, accidents, disasters, and crime) were less shared. Some topics can be shared in order to improve users' reputation. As a result, this dynamic media attention has inspired other recent studies [10].

Motivations for users to share news. There are several motivations that encourage users to share news on digital platforms. They include information seeking, socializing, entertainment, status-seeking, and prior digital platforms sharing experience on news sharing intention [140]. Based on uses and gratifications from digital platforms, and social cognitive theories, some studies show that users who are driven by the aforementioned motivations are more likely to share news on digital platforms. Personal interests [113] and prior experience with digital platforms are also significant determinants of news sharing intention [159].

Patterns of user interaction and behavior. Digital platforms have made personal contacts and relationships more visible and quantifiable than ever before [45]. Users interact by following each others' updates and passing along interesting pieces of information to their friends. This kind of word-of-mouth propagation occurs whenever a user forwards a piece of information to her friends, making users a key element in this process. Not surprisingly, a number of efforts have attempted to understand the role of users in the news ecosystem focusing on news-sharing communities (i.e., political [288]), users' subscription, and interaction patterns [8], or yet, the impact of users' beliefs and social media relationships on rumors propagation [146]. However, technologies and users are constantly evolving. Thus, studying user behavior and interaction ways is always an open issue.

User interaction by comments. In the last years, digital platforms have become social hubs for users to communicate and express their thoughts, including popular opinions or feelings toward a given piece of news [128]. Thus, writing comments has become one of the most common forms of user interaction with these collaborative systems. The interaction through comments and spaces for debates were, for a long time, widely offered in online newspapers [74]. Given this large volume of data generated from digital platforms, it became necessary to organize and summarize relevant comments [128, 154, 204] based on aspects such as quality [90], and explore strategies to distill sub-topics from all the comments related to a textual query [301], as a way to improve user experience in these environments.

Filter bubbles and echo chambers. Finally, with the insertion of digital platforms into the news ecosystem, users have a myriad of options when deciding where to get their news and what they want to read. However, due to the way news feeds and ranking algorithms work, users will only be exposed to certain news articles [108, 246]. For instance, studies show that Facebook users are connected to people with similar profiles, thus, they tend to receive news aligned with their pre-existing views [209]. This phenomenon, which favors the emergence of clusters of like-minded individuals and the polarization of opinions, is called *echo chamber*. This field has been extensively studied in recent works [86, 94, 106, 184] and as result, several other efforts have proposed ideas to mitigate the effect of *echo chambers* or *filter bubbles*, either by introducing diversity in the news that users are consuming [125, 176, 194, 218] or by highlighting posts that evoke similar reactions from opposite political views [19]. As we discuss in the following sections, the echo chamber effect may favor the spread of fake news on digital platforms.

2.2 Overview of Fake News

News media have tapped into digital platforms, and they have been a topic of interest by computer scientists. However, the changes in the news media ecosystem are still happening fast, and some of them favor campaigns of misinformation, revealing digital platforms as potential and suitable environments for spreading fake news.

Figure 2.4 shows an example of one popular fake news disseminated over digital platforms such as Facebook, Twitter, and WhatsApp about hot lemon juice being able to cure cancer. The claim was verified as "fake" by several fact-checking agencies around the world including Snopes[2], and Boatos.org[3] in Brazil. Specifically about this fake news, Snopes fact-checking agency concluded that "the best that can be said is that citrus fruits may potentially harbor anti-cancer properties that could help ward off cancer. No reputable scientific or medical studies have reported that lemons have definitively been found to be a 'proven remedy against cancers of all types', nor has any of the (conveniently unnamed) 'world's largest drug manufacturers' reported discovering that lemons are '10,000 times stronger than chemotherapy' and that their ingestion can 'destroy malignant [cancer] cells.' All of those claims are hyperbole and exaggeration not supported by facts".

Previous efforts suggest that there are at least three types of fake news [228]. The first type consists of (i) satire or parody, where sites such as the Onion[4] or Daily Mash[5] publish fake news stories as humorous attempts to satirize the media. For instance, "Jack Warner, the former FIFA vice president, has apparently been taken in by a spoof article from the satirical website The Onion"[6] after The Onion had suggested that the FIFA corruption scandal would result in a 2015 Summer Cup in the US [228]. The second type (ii) contemplates fake news that are sort of true but used in the wrong context, including hoaxes, rumors, and misleading news that are not based on facts, but supports an on-going narrative. For instance, the #Columbian Chemical plant hoax is an example of a harmful multi-platform attack[7]. Last, the third group involves (iii) news intentionally created with false information. Usually, they are fabricated and disseminated deliberately on digital platforms to either make money through the number of clicks or to cause confusion[8,9]. In this thesis we focus on exploring the type (iii).

In the next section, we present the definition of fake news adopted in this work.

[2] https://www.snopes.com/fact-check/lemon-cancer-cure/

[3] https://www.boatos.org/saude/limonada-quente-mata-cancer.html

[4] www.theonion.com

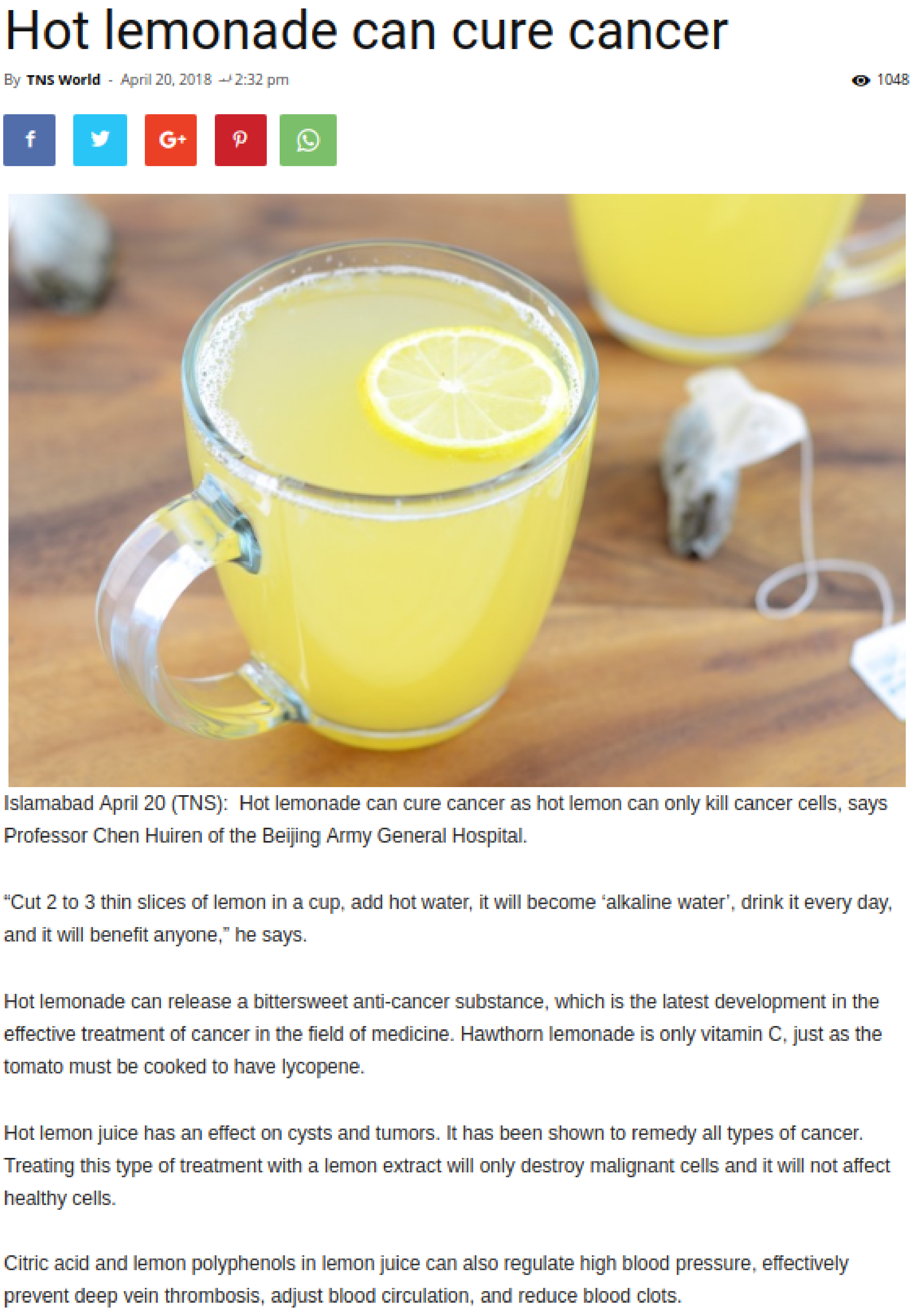
[5] www.thedailymash.co.uk

[6] https://www.theguardian.com/football/2015/may/31/ex-fifa-vice-president-jack-warner-swallows-onion-spoof

[7] https://www.nytimes.com/2015/06/07/magazine/the-agency.html

[8] http://www.cits.ucsb.edu/fake-news/danger-election

[9] https://www1.folha.uol.com.br/ilustrissima/2017/02/1859808-como-funciona-a-engrenagem-das-noticias-falsas-no-brasil.shtml



Hot lemonade can cure cancer

By **TNS World** - April 20, 2018 - 2:32 pm

1048

Islamabad April 20 (TNS): Hot lemonade can cure cancer as hot lemon can only kill cancer cells, says Professor Chen Huiren of the Beijing Army General Hospital.

"Cut 2 to 3 thin slices of lemon in a cup, add hot water, it will become 'alkaline water', drink it every day, and it will benefit anyone," he says.

Hot lemonade can release a bittersweet anti-cancer substance, which is the latest development in the effective treatment of cancer in the field of medicine. Hawthorn lemonade is only vitamin C, just as the tomato must be cooked to have lycopene.

Hot lemon juice has an effect on cysts and tumors. It has been shown to remedy all types of cancer. Treating this type of treatment with a lemon extract will only destroy malignant cells and it will not affect healthy cells.

Citric acid and lemon polyphenols in lemon juice can also regulate high blood pressure, effectively prevent deep vein thrombosis, adjust blood circulation, and reduce blood clots.

Figure 2.4: Example of fake news: Hot lemonade being able to cure cancer. Screenshot from TNS World (`tns.world/hot-lemonade-can-cure-cancer/`).

2.2.1 Fake News Definition

Fake news is a topic that still lacks a clear or universally accepted definition. According to the Collins English dictionary [52] the term "fake news" is defined as "false, often sensational, information disseminated under the guise of news reporting". However, the definition of this term (i.e., "fake news"), as well as its perception and conceptualization, has been a recent matter of debate [62, 246]. Therefore, it is crucial to state the definition we use throughout this thesis. Based on that, we define "fake news" as follows:

Definition 2.2.1. (Fake News) *"A news article or message published and propagated through media, carrying false information regardless the means and motives behind it"* [243].

2.2.2 Fake News on Digital Platforms

The general ecosystem of news, which includes fake news, has been changing over time from newsprint to radio/television and, recently, to online news and digital platforms. There are several social foundations and psychological and cognitive theories that describe the impact of fake news on both the individual and the social information ecosystem levels. First, readers prefer to receive information that confirms their existing views [185]. Second, users make choices based on the relative gains and losses as compared to their current state [268]. Finally, readers tend to believe that their perceptions of reality are the only accurate views, while others who disagree are regarded as uninformed, irrational, or biased [285]. All these factors potentiate the spread of fake news by users of digital platforms.

Although fake news detection itself is not a new problem[10], recently, some efforts are emerging aims at better comprehending the phenomenon of fake news in digital platforms [139, 161, 277]. Particularly, Vosoughi et al. [277] shows that fake news tends to spread faster than real news. Resende et al. [216] analyzed the dissemination of misinformation within WhatsApp focusing on publicly accessible political-oriented groups, collecting all shared messages during major social events in Brazil (e.g., a national truck drivers' strike and the Brazilian presidential campaign) and found the presence of fake news among the shared content using labels provided by journalists and by a proposed automatic procedure based on Google searches. Last, Lazer et al. [139] call for an interdisciplinary task force to approach this complex problem. However,

[10] `http://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535`

there are some characteristics inherent to digital platforms themselves that contribute to fake news spreading in these environments.

Malicious accounts on digital platforms. Users on digital platforms can be legitimate or not. The low cost of creating digital platforms accounts has encouraged malicious user accounts [246], such as social bots, and trolls, that are controlled by a computer algorithm to automatically interact with humans (or other bot users) on digital platforms [83]. In this context, many bots are created specifically with the purpose to do harm, such as manipulating and spreading fake news on digital platforms. There are current efforts that discuss how social bots disrupted the 2016 US presidential election online discussion [20, 26], or yet, how they coordinated disinformation campaigns during the 2017 French presidential election [81]. Recently, studies show that bots were responsible for significantly increasing the spread of fake news on digital platforms, suggesting that curbing social bots may be an effective strategy to contain the problem [242, 281].

Digital media advertising platforms. In the last years, digital media advertising platforms have evolved significantly [249]. With access to personal information and activities of millions of people around the world, these environments allow advertisers to target very specific niches of users considering personal information such as name, email address, demographic aspects, behaviors, and many others. However, targeted advertising can also be abused by malicious advertisers to efficiently reach people susceptible to false stories, stoke grievances, and incite social conflict [221]. Previous efforts highlighted several forms of abuse of targeted advertising in Facebook for inappropriately exposing users' private information to advertisers and for allowing discriminatory advertising (e.g., to exclude users belonging to a certain race or gender from receiving their ads [14, 273]). In addition, recent studies explored the extent to which political ads from the Russian Intelligence Research Agency (IRA) run prior to the 2016 US elections exploited Facebook's targeted advertising infrastructure to efficiently target ads on divisive or polarizing topics (e.g., immigration, race-based policing) at vulnerable subpopulations [221]. Overall, the results suggest that the social media ads platform can be abused by a new form of attack, which is the use of targeted advertising to create social discord reaching people susceptible to specific bits of information, including fake stories.

Echo chamber effect. Last, as mentioned previously, digital platforms have given rise to disruptive new phenomena in the news ecosystem: the so-called echo chambers[11].

[11] `https://cs181journalism2015.weebly.com/the-echo-chamber-effect.html`

Echo chambers refer to groups inside a digital platform where readers are rarely exposed to content that cuts across ideological lines but are rather fed with information that reinforces their current political or social views. While algorithmic ranking (that decides what is shown in someone's feed or search results, and in which order) can contribute to this effect, research based on Facebook data has shown that individuals' choices are the main factor in limiting exposure to cross-cutting content [22]. Recent studies have shown that the echo chamber effect facilitates access by which people consume and believe fake news due to some psychological factors: (i) relationships of the users influence their reliability in a certain source, that is, users tend to perceive a source as credible if others perceive the source as credible [246]; (ii) readers may naturally favor information they hear frequently, even if it is fake news [297], and; (iii) in echo chambers, users continue to share and consume the same information [246]. As a result, this echo chamber effect creates segmented, homogeneous, and ideologically polarized communities with a very limited information ecosystem favoring misinformation campaigns [237, 246, 256].

2.2.3 Fake News Detection

In the previous section, we introduced examples, types, definitions, and conceptual characterization of fake news on digital platforms. Now, we explore the problem definition and current approaches for fake news detection.

2.2.3.1 Problem Definition

Formally, we can define fake news detection as follows:

Definition 2.2.2. (Fake News Detection.) Given an unlabeled piece of news $a \in \mathcal{A}$, a model for fake news detection assigns a score $S(a) \in [0, 1]$ indicating the extent to which a is believed to be fake. For instance, given two unlabeled pieces of news a and a', if $S(a') > S(a)$, a' is more likely to be fake than a according to the model. A threshold τ can be defined such that the prediction function $F : \mathcal{A} \rightarrow \{\text{fake}, \text{not fake/unchecked}\}$[12] is

$$F(a) = \begin{cases} \text{fake} & \text{if } S(a) > \tau, \\ \text{not fake/unchecked} & \text{otherwise.} \end{cases}$$

[12] As we further discuss in Chapter 3, specifically for the Brazilian election dataset, that contains data from WhatsApp disseminated during the 2018 election period, we have news stories labeled as fake and news stories that have not been checked. Thus, we refer to the latter as unchecked (instead of true news), since the veracity of their content was not necessarily checked.

Essentially, fake news is a distortion bias on information manipulated by the publisher [246]. Previous efforts about media bias theory [98] show that distortion bias is usually modeled as a binary classification problem. In addition, there are related efforts that explored the detection of fake news as a binary task [53, 246, 276, 282]. Thus, based on these main reasons, we also define fake news detection in this work as a binary classification problem where the classifier's task is to distinguish fake news from others (i.e., true news and unchecked content).

2.2.3.2 Current Solutions

An effective way to detect fake news disseminated on digital platforms is the direct fact-checking, typically performed by expert journalists. A fact-checking task (i.e., the assessment of the truthfulness of a news story or claim [275]) verifies the correctness of the information by comparing them with one or more reliable sources [177]. Examples of such organizations include "Snopes.com"[13], "PolitiFact"[14], "FactCheck.org"[15], and "Aos fatos"[16], "Me engana que eu posto"[17], "e-farsas"[18], "é ou não é (G1)"[19], "Lupa"[20], "Boatos.org"[21] and "Projeto Comprova"[22], in Brazil.

However, fact-checking is a time-consuming process since it commonly requires a detailed analysis to support the verdict [275]. Consequently, traditional fact-checking cannot keep up with the enormous volume of information that is now generated online [49]. Therefore, some studies are emerging toward computational fact-checking [18, 49, 290] including automatic detection of fake news [53, 257, 276, 282].

Currently, there are mainly two approaches to perform automatically fake news detection [127]. First, (i) there are efforts that propose solutions based on artificial intelligence techniques such as supervised [53, 205, 276, 282], weakly supervised via reinforcement [284], active [27] and deep learning [130, 229, 283, 302], and also, based on specific strategies such as blockchain technology [197]. Particularly, Pérez-Rosas et al. [201] conduct a set of learning experiments to build accurate fake news detectors using sets of linguistic features. Similarly, Volkova et al. [276] build linguistic models to classify suspicious and trusted news. Typically, most of these efforts reduce the

[13] www.snopes.com
[14] www.politifact.com/
[15] www.factcheck.org/
[16] aosfatos.org
[17] veja.abril.com.br/blog/me-engana-que-eu-posto/
[18] www.e-farsas.com
[19] g1.globo.com/e-ou-nao-e/
[20] piaui.folha.uol.com.br/lupa/
[21] www.boatos.org
[22] projetocomprova.com.br/

problem to a simple classification task, in which news stories are labeled as fact/fake and a machine learning technique is then used to separate fact from fake with a model learned from the data. Specifically, these studies identify recurrent patterns on fake news after they were already disseminated to propose new features for training these models from specific data based on ideas that have not been tested in combination. Thus, it is difficult to gauge the practical potential of automatic approaches to identify fake news. As part of the research goals of this thesis, we first conducted a survey on the main features proposed in the literature for fake news detection to evaluate them in a combined way considering different scenarios. These results are further presented in Chapter 5.

Furthermore, there are recent studies aim at investigating the explainability of promising early results of the computational detection of fake news, i.e., why a particular piece of news is classified as fake [60, 156, 245, 291]. In this context, our study is complementary to previous efforts as it provides an investigation of explainable machine learning for fake news detection. However, differently from the previous studies, we conduct an in-depth investigation also exploring the explainability of fake news detection in different scenarios as a way to comprehend the structure of fake content as well as the phenomenon.

Second, (ii) other efforts to perform automatically fake news detection comprise works which aim at exploring tools or online systems for monitoring online misinformation [99]. These systems were proposed and used as countermeasures to the fake news problem on different digital platforms. Examples of such systems include Hoaxy [241], a Web platform for the tracking of social news shared containing misinformation, "Fake tweet buster" [231], a Web tool to identify users promoting fake news on Twitter, and "EleiçõesSemFake" ("Elections Without Fake")[23] in Brazil, our project to bring transparency in the dissemination of content during the 2018 Brazilian elections, as an effort to mitigate and avoid fake news dissemination.

2.2.4 Related Areas

In this section, we discuss areas related with the problem of fake news detection, highlighting some differences between them.

Rumor classification. Rumor can be defined as a story or a statement in general circulation without confirmation or certainty to facts [5]. Its main goal is to make sense of an ambiguous situation, and it can be true, false, or unverified [246]. Previous

[23] www.eleicoessemfake.dcc.ufmg.br

studies show that a rumor classification system consists of four components: rumor detection, rumor tracking, rumor stance classification, and rumor veracity classification [136, 303, 306]. When compared to fake news detection, the most related task is the rumor veracity classification, whereas stances or opinions extracted from posts in digital platforms are considered important sensors for determining the veracity of rumors [264]. In addition, these stories may include long-term rumors, such as conspiracy theories, as well as short-term emerging rumors. Unlike them, fake news refers to information related specifically to public news events that can be verified as fake [246].

Truth discovery/credibility. The problem of detecting true facts from multiple conflicting sources is called truth discovery [147], which aims to determine the source credibility and object truthfulness at the same time [246]. Some efforts show that there are measurable differences in the way credible and not credible messages propagate on Twitter [38]. Although most existing truth discovery approaches focus on handling structured input, while social media data is highly unstructured and noisy, the fake news detection problem can benefit from various aspects of truth discovery approaches such as, for example, the features considered. However, truth discovery methods cannot be well applied when a fake news article was just released and published by only few news outlets because, at that point, there are not enough digital platforms posts relevant to it to serve as additional sources [246].

Clickbait detection. *Clickbait* refers to "content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page" [191]. Studies show that such mechanism contributes to the spreading of fake news on the Internet [47]. Thus, there are current works that aim to detect and prevent clickbaits in online news media [44, 138]. Even though not all fake news may include clickbait headlines, specific clickbait headlines can be useful to detect fake news [246].

Spammer and bot detection. Evidences suggest that malicious accounts (i.e., bots, trolls, etc) are key elements for spreading fake news on digital platforms [242], as aforementioned. Different approaches to understanding and detecting malicious user behavior have been extensively explored in previous efforts [24, 25, 271]. In summary, bots[24] can increase the circulation of false information on digital platforms by giving a false impression that information is highly popular and endorsed by many people, which enables the echo chamber effect to bolster the propagation of fake news [242]. Hence, both spammer and social bots could provide valuable insights about target specific

[24] Bots (Short for software robots), are accounts controlled by software, algorithmically generating content and establishing interactions [271].

malicious digital platforms accounts that can be used for fake news detection [246].

2.3 Research Gaps

In spite of several efforts that explore solutions to automatic fake news detection, there is still space for advancements. Next, we describe the research gaps that guided our work:

1. Most of the existing efforts for fake news detection are either concurrent works, which identify recurrent patterns on fake news after they have been already disseminated, or approaches that propose new features for training classifiers based on ideas that have not been tested in combination. Thus, this work surveys existing studies on this topic, identifying the main features proposed for this task. We implement these features, propose new ones, and test the effectiveness of a variety of supervised learning classifiers when distinguishing fake from real stories. In addition to exploring recently released and fully-labeled data from different scenarios, we build a new dataset of fact-checked news stories shared in WhatsApp during the 2018 Brazilian elections which can be useful for research in a variety of contexts.

2. Previous efforts proposed single models as an effective way to solve the fake news problem on digital platforms. We have seen that fake news may have different characteristics (e.g., subjects, sources with a distinct bias and styles, etc). Thus, we believe there is no single model to tackle all facets of fake news detection. In addition, little is known about the discriminating power of features proposed in the literature, either individually or when combined with others. Some may be adequate for pinpointing fake news with specific patterns, while others are more general but not sufficiently discriminating. Moreover, while explaining the decisions made by proposed models is central to understand the structure of fake content, this discussion is often left aside. In this work, we address all these issues.

3. We do not find efforts that explore automatic fake news detection focusing on investigating whether there is a set of features that remain useful to build models with high performance which are able to identify fake news disseminated on digital platforms considering different scenarios. In this thesis, we explore data from the 2016 US and 2018 Brazilian elections.

4. Last, although there are some isolated initiatives that study fake news spread in Brazil [172], to the best of our knowledge, this is the first work that explores strategies with practical potential for detecting fake news spread on WhatsApp. Therefore, we propose and implement a new ranking mechanism that accounts for the potential occurrence of fake news within the data, significantly reducing the number of content journalists and fact-checkers have to go through before finding a fake story. Then, we deploy our approach in a real system, the WhatsApp Monitor proposed as part of the "EleiçõesSemFake" project introduced in this chapter.

2.4 Summary

In this chapter, we reviewed related works and the background knowledge required to understand this thesis, highlighting changes in the news ecosystem from the rise of digital platforms to how computer scientists are studying 'news' under this new scenario. In addition to defining fake news, we presented current approaches proposed to tackle this problem. We also highlighted some factors of digital platforms that drive the propagation of fake news. Last, we described some related areas to fake news detection and presented the research gaps addressed in this thesis. The next chapter describes the datasets for fake news detection available in the literature, including those used in this work, and our effort to build a new dataset containing news stories disseminated during the 2018 Brazilian presidential election.

Chapter 3

Datasets

To make concrete contributions towards fake news understanding and detection, researchers need a wide and broad set of data containing labeled instances, i.e., fact-checked content, covering different topics and contexts [118, 187]. In the particular context of fake news during election campaigns, data covering multiple elections is also of interest as it can unveil potentially different properties or reinforce common characteristics.

Therefore, in this thesis, we first perform a brief survey on existing public datasets commonly used by concurrent works investigating the phenomenon of fake news, either to understand it or to propose solutions that aim to minimize the effects caused by it. Those datasets often label content as fake or true stories. This fact-checked content can appear in different formats, such as news articles, claims, or quotes from celebrities, rumors, reports, or images, and for different scenarios such as elections, wars, and health.

Table 3.1 summarizes some of the well-known fact-checked datasets and their main characteristics, including a description, the total number of instances as well as their context distribution by label (i.e., fact-checking verdict - true; fake; and its variation) and information about raters (i.e., fact-checkers)[1]. Note that we colored in red the number of fact-checked instances labeled as fake, in blue, the true news ones, and in black the remaining ones (i.e., those that are neutral, no factual, etc).

[1] Table 3.1 does not include datasets which are out of the scope of this work such as rumors [135], stance detection (`https://github.com/FakeNewsChallenge/fnc-1`) [84], and credibility [171].

Table 3.1: Labeled datasets for fake news detection task.

Dataset	Description	Topic	Labels	Raters	# Instances
BuzzFace[2] [202, 235]	News published in Facebook from 9 agencies over a week right before the occasion of the 2016 US election.	Elections	mostly false (104), mixture of true and false (245), mostly true (1,669), no factual (264)	Journalist experts from BuzzFeed.	2,282
CoAID[3] [59]	News and claims related to COVID-19 on websites and social platforms, along with users' social engagement about such news.	Health	fake (162), true (1,734)	Reliable media outlets and fact-checking websites.	1,896
Fact-Checked-Stat [275]	Statements fact-checked from popular fact-checking websites labeled by journalists.	General	true (32), mostly true (34) , half true (68), mostly false (37), false (49), fiction (1)	Journalists from fact-checking websites.	221
FakeHealth [63]	A repository that consists of two datasets, i.e., Health-Story and HealthRelease and includes news contents about health, news reviews, social engagements, and user networks.	Health	fake (763), true (1,533)	Expert reviewers in the health domain.	2,296
FA-KES [232]	A fake news dataset around the Syrian war (i.e., reports on war incidents that took place from 2011 to 2018.)	War	fake (378), true (426)	Semi-supervised fact-checking labeling approach.	804
Fake.Br Corpus [172]	True and fake news that were manually aligned, focusing only on Brazilian Portuguese.	General	fake (3,600), true (3,600)	Researchers.	7,200
Fake-News-Net[4] [246, 247]	A repository for an ongoing data collection project for fake news research including news content and social context features with reliable group truth fake news labels.	General	fake (211), real (211)	Journalists experts from BuzzFeed and fact-checkers from PolitiFact.com.	422

Continue on next page

[2] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

[3] https://github.com/cuilimeng/CoAID

[4] https://github.com/KaiDMML/FakeNewsNet

Dataset	Description	Topic	Labels	Raters	# Instances
Fake-Real-News[5]	News articles published during 2015-2016 along with their titles. The entire corpus was built crawling real news with New York Times and NPR APIs[6] and fake news from Kaggle[7] dataset items to ensure an uniform distribution of the samples from both the classes.	General	fake (3,164), real (3,171)	Journalists for true news and human annotators from BS Detector for fake news.	6,335
Fake-Satire [101]	Dataset of fake news and satire stories that are hand-coded, verified, and, in the case of fake news, include rebutting stories.	General	fake news (283), satire (203)	Researchers based on an article from a fact-checking site or a piece of information that disproves a claim.	486
Fake-Twitter-Science [277]	All of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ∼126,000 stories (rumors cascades) tweeted by ∼3 million people more than 4.5 million times.	General	true (24,409), false (82,605), mixed (19,287)	Agreement between fact-checkers from six independent fact-checking organizations.	126,301
Kaggle[8]	Text and metadata from fake and biased news sources around the web from BS Detector.	General	bias (443), bs (11,492), conspiracy (430), fake (19), hate (246), junksci (102), satire (146), state (121)	Human annotators from BS Detector.	12,997
LIAR [282]	Short statements from PolitiFact.com manually labeled.	General	half-true (2,638), false (2,511), mostly-true (2,466), barely-true (2,108), true (2,063), pants-fire (1,050)	Fact-checkers from PolitiFact.com.	12,836
NELA-GT-*[9] [103, 187]	News articles from various news and media outlets including mainstream, hyperpartisan, and conspiracy sources.	General	unreliable, mixed, reliable	Source-level ground truth labels from 7 different assessment sites.	1.12M

[5] https://github.com/GeorgeMcIntire/fake_real_news_dataset

[6] https://developer.nytimes.com/ and https://www.npr.org/api

[7] www.kaggle.com

[8] https://github.com/JasonKessler/fakeout

[9] https://doi.org/10.7910/DVN/O7FWPO

At a high-level, these datasets of fact-checked content were labeled according to different scales, as either fake or true by expert journalists, fact-checking websites, and industry detectors[10], providing different pieces of information and contexts that allow us to extract distinct types of features [246].

As introduced in Chapter 1, the research goal of this thesis includes investigating the ability of current supervised machine learning approaches to identify fake news considering data from two major political events that have been extensively abused by misinformation campaigns [26, 30, 216] (i.e., the 2016 US and 2018 Brazilian elections). Furthermore, it is desirable to measure the potential of our features to detect fake news in a scenario different from politics (i.e., health). Thus, to accomplish this goal, we need datasets covering these scenarios. Based on our survey on existing datasets for fake news detection listed in Table 3.1, there are no available datasets covering all scenarios of interest (e.g., we did not find available data from the 2018 Brazilian elections). Also, it is necessary that the selected datasets contain for each news story labeled by specialists, their textual content, information about their sources, and about the dissemination of these news, particularly in social platforms. This enables the implementation of all features for fake news detection from previous efforts and eventually, the proposition of new ones. Next, we present details of each of the selected datasets covering each of the scenarios explored in this work: the 2016 US election, the 2018 Brazilian election, and Health.

3.1 2016 US Election

We use the BuzzFace dataset [235]. It contains 2,282 news articles labeled by BuzzFeed journalists related to the 2016 US election [250]. The BuzzFace dataset consists of an enriched version of the one created by BuzzFeed, with over 1.6 million comments associated to the news stories as well as shares and reactions from Facebook users. The news stories in the dataset are labeled into four categories: mostly true, accounting to 73% of all news articles, mostly false (4%), mixture of true and false (11%), and non-factual (12%). For simplicity, we discarded the non-factual content[11] and merged

[10] BS Detector: `https://gitlab.com/bs-detector/bs-detector`

[11] **Note:** A typical pre-processing step is to separate factual from non-factual content. This task is easier than classifying factual data as fake or true since it is not necessary to check the veracity of the information using external sources. For illustration purposes, we conduct a small experiment to evaluate the accuracy of XGB [46] when discriminating factual and non-factual news using the features that will be described in Chapter 4. Our simple classifier performed very well, yielding 0.882±0.024 of AUC. It is possible to achieve even higher performance levels by choosing features better tailored for this task. For this reason, this work assumes that non-factual data was already removed and only factual data is used as input. The alternative approach is to consider a multi-label classification

the mostly false with the mixture of true and false into one single class, referred as "fake news" (349 out of $2,018$ stories). The rationale is that stories that mix true and false facts may represent attempts to mislead readers. Thus, we focus our analysis on the understanding of how features are able to distinguish two classes, true and fake news.

We select and use this dataset for the following reasons: (i) it has a good trade-off between the volume of instances and the variety of information provided (i.e., textual content, information about sources, and about the dissemination of these news, specifically in digital platforms) and; (ii) particularly, we believe that this dataset best represents the real world, where the proportion of fake news fabricated and shared is smaller compared to the large volume of news generated every day.

3.2 2018 Brazilian Election: A New Dataset

As previously mentioned, we does not find available data from the 2018 Brazilian elections (see Table 3.1). Thus, we build a new dataset of fact-checked news stories disseminated in Brazil during the 2018 election period. Specifically, we collected news stories published on WhatsApp during the period of interest. This platform was chosen for the following reason: the WhatsApp was pointed as one of the main vectors of fake news spreading during the 2018 presidential Brazilian elections [258] and source for distributing political messages in bulk[12]. Also, a recent report from the Reuters Institute found that in countries such as Brazil the WhatsApp has become a primary network for discussing and sharing news [183]. Next, we describe how we built our dataset, presenting its properties and limitations.

3.2.1 Dataset Construction

To gather fact-checked news stories shared on WhatsApp during the 2018 Brazilian election, the first step is collecting the data from WhatsApp, which is not a trivial task. WhatsApp is an encrypted and very closed network, from which is hard to extract data. Thus, to collect data from WhatsApp, we followed the approach used by [95] and [216] to get access to messages posted on public WhatsApp groups[13]. Their approach joins a set of public groups on WhatsApp and collects all messages (including text, images,

problem, but this has the potential to increase the number of instances that need to be verified by an expert.

[12] https://www.bbc.com/news/technology-45956557

[13] Whatsapp groups are made effectively publicly accessible when group administrators openly share invitation links on the Web and online digital platforms.

audios, and videos) shared on the application. Figure 3.1 presents an overview of our data collection process.

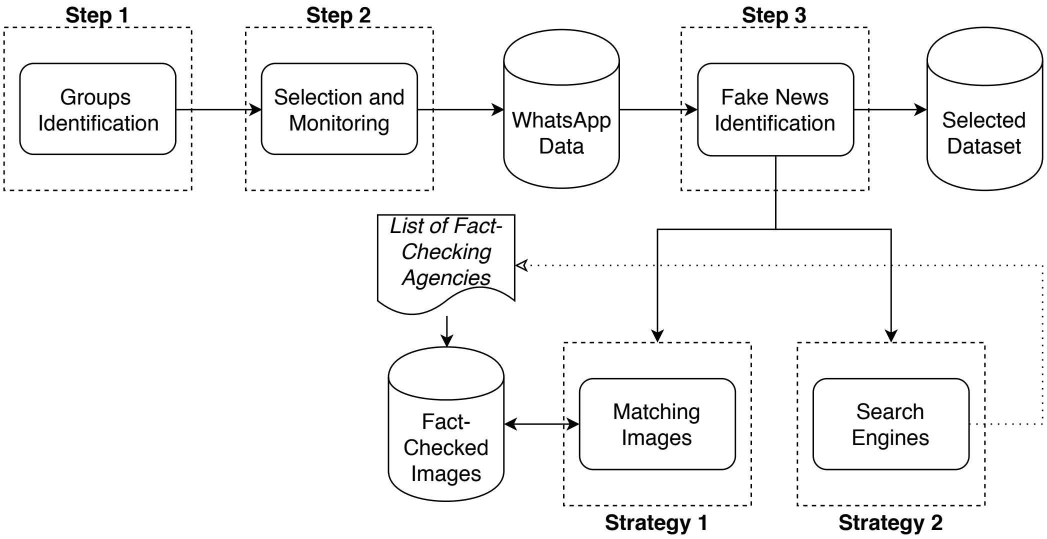


Figure 3.1: Overview of data collection.

Given a set of invitation links to public groups (**Step 1**), we automatically joined these groups and saved all data coming from them. We selected over 400 Brazilian groups dedicated to political discussions which we monitored during the election period, i.e., August-November 2018.

In addition to exploring these groups which are quite widely used in Brazil [155, 183], we choose to filter only messages that disseminated news stories through images. Previous efforts showed that images are the most frequent type of media content, as well as an important source of fake news [216]. Also, images are harder to manipulate and can be easily shared across groups and even platforms, unlike text, which is much easier to change during dissemination. Thus, we monitored the groups (**Step 2**) and, for each message collected, we filter out those containing images. This leads us to a set of over 34K images shared by more than 17K users in Brazil from which we extracted the following fields: (i) ID of the group the message was posted, (ii) user ID[14], (iii) timestamp, and (iv) the attached multimedia files (i.e., images). We emphasize that all sensitive information (i.e., group names and phone numbers) were anonymized in order to ensure the privacy of users.

In the next step, the images were clustered based on their similarity using perceptual hashing techniques to group together all images that are visually similar [299]. This allows us to track the spread of an image through multiple groups on WhatsApp,

[14] Note that we consider that these are, actually, phone numbers. Hence, different numbers by the same person are different identifiers.

including the metadata on which users and groups shared that image and the time of sharing. This is valuable information that not only shows the popularity of the images but also helps in the analysis of dissemination and reach of images within the WhatsApp. Furthermore, it also aids us in the next task to create a set of unique images that need to be checked to determine whether they contain fake news stories.

3.2.1.1 Fake News Identification

After tracking all images shared on WhatsApp and their dissemination from selected public groups, we next identify which images contain fake news stories. This step (**Step 3**) consists of identifying, among the images that circulated in WhatsApp during the monitored period, those that disseminated fake news by matching the images from WhatsApp and their correspondent occurrence images that were already fact-checked by major online fact-checking agencies in Brazil. We accomplish this task through two distinct approaches: (i) based on the match of fact-checked images, and (ii) by using search engines. We use both as complementary strategies to increase the amount of data found. Next, we describe the details of our proposed strategies.

Matching with Fact-Checked Data. First, we crawled all images which were fact-checked from popular fact-checking websites in Brazil[15]. For each that fact-checking website, we developed a script able to parse and save all content checked, as well as all images on same webpage related to the content checked and, when explicitly available, the label is given for that checking. Note that not all images contain fake news stories as many fact-checking agencies do not have the explicit label. We only used these images as a noisy source of ground truth. Therefore, for this work, we call fact-checked images all different images files related to fact-checked content collected from fact-checking agencies by this methodology. In total, we collected over 100k fact-checked images from Brazil.

After that, we match this set of fact-checked images and those circulating on WhatsApp using a perceptual hashing approach that generates hashes to compare visually similar images. We used the state-of-the-art technique developed and being used at Facebook: the PDQ hashing[16]. This approach, an improvement over the commonly and broadly used pHash [299], can detect near visually similar images even if cropped differently or have small amounts of text overlaid on them. Facebook PDQ hash produces a 256 bit hash string using a discrete cosine transformation

[15] aosfatos.org, veja.abril.com.br/blog/me-engana-que-eu-posto/, www.e-farsas.com, g1.globo.com/e-ou-nao-e/, piaui.folha.uol.com.br/lupa/, and www.boatos.org

[16] https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf

Table 3.2: WhatsApp collection – fact-checked images shared in WhatsApp during the 2018 Brazilian election.

	#Users	**#Groups**	**News Stories (Unique Images)**	**#Fake News**	**Time Span**
Brazil	17,465	414	4,524	135	2018/08 - 2018/10

algorithm. We used the PDQ algorithm to compute the pairwise match between each fact-checked image collected and each image shared in our WhatsApp dataset. For all instances matched this way, we manually verified if the image was labeled as fake by the fact-checking agency.

Search Engines. In our second strategy, we automated the process of searching each image shared on the WhatsApp groups on the Web by using the Reverse Google Image search as proposed in [216]. Given the search results for an image, we checked whether any of the returned pages belongs to one of the main fact-checking domains from Brazil, according to the previously defined set. If so, we parsed the fact-checking page and automatically labeled the fact-checked image as fake or true depending on how the image was tagged on the fact-checking page.

As shown in Table 3.2 our final dataset is composed of more than 4.5k distinct images shared by more than 17k users in 414 different groups. Also, it contains 135 distinct fact-checked images from Brazil. Specifically for the Brazilian election dataset, we have news stories labeled as fake and news stories that have not been checked. Thus, in the rest of the thesis, we refer to the latter as unchecked, since the veracity of their content was not necessarily checked. Thus, we do not claim that there is no fake news in the unchecked content, given that such an assertion is restricted to the availability of checked facts.

3.3 Health

Last, we use the FakeHealth dataset[17] that was built based on two repositories: HealthStory and HealthRelease. Each repository includes health news stories comprising their content (e.g., text, source, image), reviews, social engagements (e.g., number of replies and retweets on Twitter), and user networks [63].

The news stories in this dataset were evaluated on the HealthNewsReview.org[18], a web-based project which analyzes news stories about health care interventions to

[17] https://doi.org/10.5281/zenodo.3606757

[18] https://www.healthnewsreview.org/

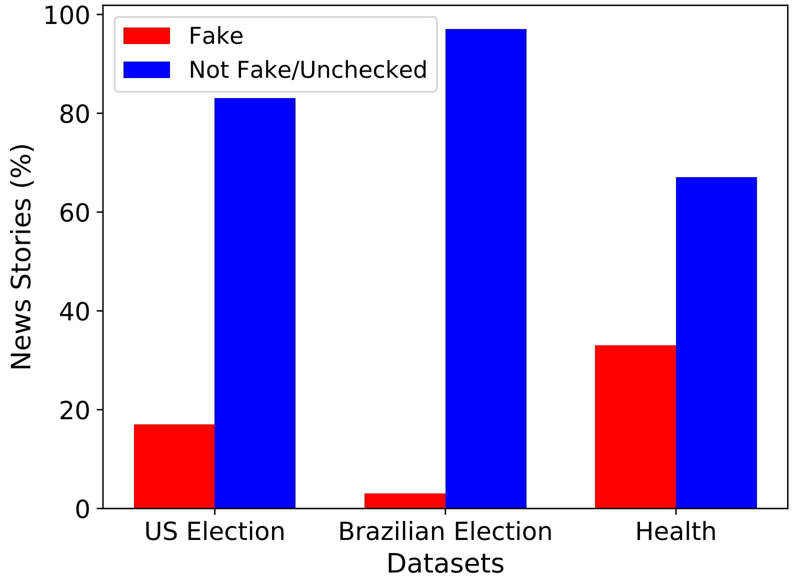


Figure 3.2: Summary of labeled instances in each dataset.

improve the quality of healthcare information, including particular treatments, tests, products or procedures. Specifically, each news was checked by reviewers with years of experience in the health domain based on some criteria. These criteria assess the health news in diverse aspects such as the overclaiming, missing of information, reliability of sources, and conflict of interests [63]. The final dataset contains 763 and 1,533 news labeled as fake and real respectively, totalizing 2,296 news health stories.

3.4 Summary

We started this chapter by surveying existing open datasets for fake news detection. Then, we presented the selected datasets in this work covering each of the scenarios of interest: the US and Brazilian elections, and Health which contains different proportions of labeled as fake or true/unchecked stories, as shown in Figure 3.2. Furthermore, we described how we collected fake news stories disseminated during the 2018 Brazilian election releasing a novel dataset to the research community.

We need to acknowledge that there are some limitations regarding the data gathered in this thesis. First, we used two different strategies in order to detect the highest amount of fact-checked images shared on WhatsApp, but it is possible that many existing fake news from our dataset were not checked by any of the fact-checking agencies used in this thesis or even images that did not match using the hashing technique. Thus, we cannot comment on the recall of our dataset and the sampling bias present in the dataset. Moreover, the WhatsApp public groups used here are just a portion of the entire network of WhatsApp and may not be statistically representative, though we cannot have access to all groups on WhatsApp to perform a real analysis. Still, to date, this is the largest available sample of WhatsApp for research.

The 2018 Brazilian election dataset presents also some strong advantages. First, it uses labels from fact-checking agencies, relying on specialist labeling. Also, a dataset based on news stories disseminated through images is less common than a text-based one on fake news as we see on our survey presented in this chapter. Furthermore, it covers an important aspect of studies on fake news around the world: the elections period. On the other hand, it is restricted to peculiarities of only one isolated event. Finally, the dataset explores the context of the closed network of WhatsApp in which shared content is usually popular in other media too. WhatsApp is becoming very important to studies on fake news, especially in Brazil, the country from which we gather data. However, it is a challenging task to get any data from this encrypted messaging app due to the closed nature of its network. Thus, this dataset can provide a useful resource to shed light on this phenomenon, principally in terms of spreading as we include information about dissemination.

The next chapter describes our survey on existing features for fake news detection and our effort in the direction of implementing these features using the selected datasets in this chapter, including the new ones proposed in this thesis.

Chapter 4

Features for Fake News Detection

Fake news detection on traditional news media mainly relies on news content, while in digital platforms, we can use side information (e.g., the number of shares, comments, etc) as additional information to help detect fake news [246]. The literature that explores features for fake news detection is quite broad considering efforts related to information credibility, rumor detection, and news spread. Hence, we conduct a systematic survey on these efforts, aiming to identify proposed features. Table 4.1 presents a summary of this survey along with some of the techniques used to extract those features.

In summary, we can categorize features explored in previous works as follows: (i) features extracted from content (e.g., language processing techniques and image properties) [107, 114, 276, 286, 303]; (ii) features from source (e.g., reliability and trustworthiness) [148]; and finally (iii) features extracted from the environment, which usually involves propagation dynamics within digital platforms and Web [49]. Next, we describe how we implemented or adapted the features summarized in Table 4.1.

Table 4.1: Overview of features for fake news detection presented in previous work.

Extracted from...	**Feature Set**	**Techniques mostly used**	**References**
Content	Language Structures (Syntax)	Sentence-level features, such bag-of-words approaches, “n-grams", part-of-speech (POS tagging)	[53, 135, 201, 212, 227, 246, 286]
	Lexical Features	Character level and word-level features, such as number of words, characters per word, hashtags, similarity between words, etc	[3, 27, 38, 107, 114, 131, 201, 202, 212, 222, 246, 286, 303]
	Moral Foundation Cues	Moral foundation features	[276]
	Images and Videos	Indicators of manipulation and image distributions	[117, 121]
	Psycholinguistic Cues	Additional signals of persuasive language such as anger, sadness, etc and indicators of biased language	[107, 135, 201, 227, 276, 277]
	Semantic Structure	Word embeddings, “n-grams" extensions, topic models (e.g., latent Dirichlet allocation (LDA)), contextual informations	[27, 49, 53, 88, 227, 282, 286, 303]
	Subjectivity Cues	Subjectivity score, sentiment analysis, opinion lexicons	[212, 227, 234, 276]
Source	Bias Cues	Indicators of bias (e.g., politics), polarization	[119, 219, 222]
	Credibility and Trustworthiness	Estimation of user’ perception of source credibility	[38, 242, 246]
Environment (Digital Platforms and Web)	Engagement	Number of page views, likes (on Facebook), retweets (on Twitter), etc	[38, 85, 88, 107, 131, 241, 246, 257, 277]
	Network Structure	Friendship network, complex network metrics (i.e., degree, average shortest path)	[38, 53, 88, 107, 131, 212, 241, 242, 246, 248, 267, 276, 277]
	Temporal Patterns and Novelty	Time-series, propagation, novelty metrics	[38, 85, 88, 135, 241, 242, 246, 267, 277]
	User Information	Users’ profiles and characteristics across individual level and group level (e.g., their friends and followers)	[38, 107, 135, 212, 222, 241, 242, 246, 257, 267, 277]

4.1 Our Implementation of Features for Fake News Detection

As previously mentioned, most of the existing efforts to detect fake news propose features that leverage information present in a specific dataset. In contrast, some of the datasets used in this work allow implementing most of the features explored in previous works. Next, we briefly describe how we implemented or adapted the features summarized and detailed in Table 4.1. In total, we implemented 208 features for fake news detection. In addition to features typically used for this task, we propose 22 new features for fake news detection. Specifically, we propose features based on image properties, semantic structure, news sources (i.e., specific attributes of publishers), and new propagation measures within/outside digital platforms. Table 4.2 presents an overview of our implementation of features for fake news detection.

4.1.1 Content Features

Content features involve not only the news story but also its headline, associated images and any message that was published along with it. Thus, specially for news stories embedded in images and videos[1], we apply image processing techniques, e.g., optical character recognition (OCR) provided by Google Vision API[2], for extracting the text shown on them as a way to implement textual features. In total, we evaluated 154 content features. The sets of features are described next.

Image properties (IMAG). The images are part of the news and provide visual cues to frame the story [246]. Thus, we also use the Google Vision API to extract various pieces of information associated with images such as number of labels, colors, objects, and the presence of faces. Further, we use as features measures provided by safe search detection, which detects explicit content (adult, spoof, medical, violence, and racy) and returns the likelihood that each is present. In total, we implemented 10 image features.

Language structures (SYNT, for Syntax). Sentence-level features, including bag-of-words approaches, "n-grams" and part-of-speech (POS tagging) were extensively explored in previous efforts as features for fake news detection [53, 246]. Here we implement 31 features from this set including number of words and syllables per sen-

[1] Regarding the Brazilian election dataset, i.e., WhatsApp data, we consider only news stories disseminated through images. Previous efforts showed that images are the most frequent type of media content, as well as an important source of fake news [216].

[2] `https://cloud.google.com/vision`

Table 4.2: Our implementation of features for fake news detection.

Content (154 Features)	IMAG (10)	Number of faces in image†, labels†, colors, objects, Web entities* and safe search indicators* (adult, spoof, medical, violence, and racy)
	SYNT (31)	readability indicators* (gunning fog [105], linsear write formula, smog index [163], flesch kincaid grade index [251], difficult words, LIX [13], dale chall readability score [89], RIX, coleman liau index [51], automated readability index (ARI) [239], flesch reading ease index [87]), sentence begin (article, interrogative, preposition, pronoun, conjunction, subordination), sentence information (summary, complex words, characters, paragraphs, long and short sentence, word types, words per sentence, syllables, type token ratio)
	LEXI (59)	word count, count-low-word*, count-up-word*, words per sentence, summary variables (analytic thinking, clout, authentic, emotional tone), language metrics (words > 6 letters, dictionary words from LIWC), function words (function, total pronouns, personal pronouns – 1st pers singular, 1st pers plural, 2nd person, 3rd person singular, 3rd person plural – impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, negations), word usage (regular verbs, to be verb, adjectives, comparatives, interrogatives, numbers, quatifiers, nominalization), informal speech indicators (swear words, netspeak, assent, nonfluencies, fillers), punctuations (periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses – pairs –, others punctuations), occurrence or not of specific characters (i.e.,#)
	PSYC (44)	affective indicators (affect, anxiety, anger, sadness), social and personal concerns (social, family, friends, female and male referents, biological processes, body, health/illness, sexuality, ingesting, drives, affiliation, achievement, power, reward focus, risk focus, work, leisure, home, money, relig, death), cognitive indicators (cognitive process, insight, cause, discrepancies, tentativeness, certainty, differentiation, perception, seeing, hearing, feeling) and temporal references (past focus, present focus, future focus, relative, motion, space, time)
	SEMA (5)	toxicity*, category, topic, contextual information, labels
	SUBJ (4)	subjectivity, sentiment, positive and negative emotions
	TYPE (1)	type of content (e.g., video, image, link)

Continue on next page

Source (13 Features)	PUBL (7)	publisher, page/group that published news story, ip (ASN)⋆, latitude⋆, longitude⋆, the user who carried out the first news share⋆, the groups where this news story was disseminated⋆
	CRED (5)	page talking about count⋆, page fan count⋆, indicators of low credibility, relative position of news domain on the Alexa ranking⋆, dissimilarity between domains from the Alexa Ranking and news domains⋆
	BIAS (1)	political bias
Environment (41 Features)	ENGA (18)	counts of social interactions (e.g., reactions, shares, and comments in Facebook, retweets and replies in Twitter, shares in WhatsApp)‡
	EXPR (10)	external propagation measures†, i.e., information about the dissemination of the news stories on the Web⋆
	TEMP (13)	rate at which shares/comments are made‡

⋆ New features.

‡ We consider intervals time: 900, 1800, 2700, 3600, 7200, 14400, 28800, 57600, 86400, 172800, 259200, 345600, and 432000 seconds.

† Indicates features not previously used for fake news detection.

tence as well as tags of word categories (such as noun, verb, adjective). In addition, to evaluate writers' style as potential indicators of text quality, we also implement features based on text readability which is concerned with how (unnecessarily) complex is the writing in terms of word and grammar usage [64].

Lexical features (LEXI). Typical lexical features include character and word-level signals [38, 246], such as amount of unique words and their frequency in the text. We implement 59 lexical features, including number of words, first-person pronouns, demonstrative pronouns, verbs, hashtags, all punctuations counts, etc.

Psycholinguistic cues (PSYC). Linguistic Inquiry and Word Count (LIWC) [199, 261] is a dictionary-based text mining software that categorizes words into psychologically meaningful groups. Since its proposition, it has been widely used for a number of different tasks, including sentiment analysis [220] and discourse characterization in digital platforms [56, 274]. Thus, we use the dictionaries for the English and Portuguese languages, which are organized as a hierarchy of LIWC categories [198]. For example, the word 'cried' falls into the sadness, negative emotion, overall affect, verbs, and past focus categories. Examples of words representing *Religion* in the dictionary are *altar, church*. Then, in a given text, the LIWC software finds the occurrence of the words in each category. The output is the proportion of the words in each category to the total words in the text. We use its latest version (2015) to extract 44 features that capture additional signals of persuasive and biased language.

Semantic structure (SEMA). There are features that capture the semantic aspects of a text [53, 282] that are useful to infer patterns of meaning from data [36]. As part of this set of features, we consider category (i.e., sport, health, etc) and topic (i.e., elections, etc) of news article[3] and the toxicity score obtained from Google's API[4], called Perspective. The API uses machine learning models to quantify the extent to which a text (or news headline, for instance) can be perceived as "toxic". Furthermore, based on the images associated with news stories, we also use the Google Vision API to extract contextual information of them such as topics/labels.

Subjectivity cues (SUBJ). Last, previous efforts show that the sentiment of the news article is strongly related to the popularity of the news and also with the dynamics

[3] This is important because although we are analyzing specific scenarios, news stories can be associated with more than one category/topic.

[4] `https://www.perspectiveapi.com/#/`

of the posted comments on that particular news [213]. Hence, we use a Portuguese and English versions[5] of SentiStrength method [263] to measure the polarity of each news story. SentiStrength is a well-established method that implements a combination of supervised learning techniques with a set of rules that impact the "strength" of the opinion contained in the text in a scale from -5 (very negative) to +5 (very positive). Also, using TextBlob's API[6], we compute the subjectivity score of a text.

4.1.2 Source Features

Source features are related to the publisher of the news story, e.g., domain information that can be extracted from news URLs, indicators of the credibility and political bias, etc. Thus, we extract features of this set through two distinct approaches: for the cases where there is a URL associate with the news story (e.g., the US election and Health datasets), (i) we first parse all news URLs and extract the domain information. When the URL is unavailable, we associate the official URL of news outlet to the news article. In this scenario, we extract indicators of credibility and source trustworthiness. Apropos, we introduce a new set composed by 5 (five) features, called domain/publisher as we further discuss. Regarding the Brazilian election dataset, where there is no URL associated with the news story, (ii) we use as publishers the users and groups within WhatsApp to capture information about the potential fabricators of news stories.

Domain (PUBL, for publisher). Ever since creating fake news became a profitable job, some cities have become famous because of residents who create and disseminate fake news[7,8]. In order to exploit the information that publisher localization could carry, two distinct pipelines were built: (i) for the cases where there is a URL associated with the news story (i.e., the US Election and Health datasets), we take each news website URL and extract new features, such as IP address, latitude, and longitude. First, for each domain, the corresponding IP address is extracted using the traceroute Linux command, which prints the route that a packet takes to reach the host. Then the ipstack API[9] is used to retrieve the location features. Although localization information (i.e., IP address) has previously been used in works that exploit bots or spam detection [211], to the best of our knowledge, there are

[5] `sentistrength.wlv.ac.uk`

[6] `http://textblob.readthedocs.io/en/dev/`

[7] `https://www.bbc.com/news/magazine-38168281`

[8] `https://www1.folha.uol.com.br/ilustrissima/2017/02/1859808-como-funciona-a-engrenagem-das-noticias-falsas-no-brasil.shtml`

[9] `https://ipstack.com/`

no works that exploit this data in the context of fake news detection context[10]. Nonetheless, (ii) for the cases where there is no URL associetad with the news story (i.e., Brazilian election dataset), first, we consider the anonymized identifier of the user who shared a news story for the first time as a categorical feature. This identifier incorporates the user's locale code. Similarly, we capture the first WhatsApp group in which it was posted. In a preliminary analysis, we found that only 10 users were responsible for the first post of 23% of images fact-checked as fake, and that 44% of these images had their first appearance in just 9 distinct groups. We conjecture that these publisher dynamics provide valuable information, capable of capturing any indication of a malicious and orchestrated action to intentionally spread fake news.

Credibility and trustworthiness (CRED). In this feature set, we introduce 5 (five) new features to capture aspects of credibility (or popularity) and trustworthiness of domains[11]. Using Facebook's API[12], we collect user engagement metrics of Facebook pages that published news stories (i.e. page talking about count and page fan count). Then, we use the Alexa API to get the relative position of news domain on the Alexa Ranking[13]. Furthermore, using this same API, we collect Alexa's top 500 newspapers. Based on the hypothesis that some unreliable domains may try to disguise themselves using domains similar to those of well-known newspapers, we define the dissimilarity between domains from the Alexa ranking and news domains in our dataset (measured by the minimum non-zero edit distance) as features. Last, we use indicators of low credibility of domains compiled in [242] as features.

Bias cues (BIAS). The correlation between political polarization and spread of fake news was explored in previous studies [219, 222]. In this thesis, for the US Election dataset (i.e., BuzzFace), we use the political biases of news outlets as a single feature. For the Health dataset, this feature has not been implemented. Last, for the Brazilian election dataset, we infer the political biases of publishers WhatsApp groups according to the following strategy: (i) we automatically parse the group description (i.e., group name) to check whether there is any information about its political bias. If so, based on [67], we automatically label the group as "right", "left" or "mainstream". For example, the group "#BOLSONARO PRESIDENTE" was assigned "right" as political

[10] We conduct a IP to ASN mapping using `https://asn.cymru.com/cgi-bin/whois.cgi.` and exploit it as categorial feature.

[11] This group of features was implemented only for cases where there is a URL / domain associated with a news story, i.e. the US election and Health datasets.

[12] `https://developers.facebook.com`

[13] `https://www.alexa.com`

bias since Jair Bolsonaro, back then a presidential candidate, is a right-wing partisan. For cases where the description of the group does not provide any indication of its political alignment, (ii) we manually inspect the group content, that is, the bias of messages posted in them, in order to infer the political bias of group. This strategy has been used in previous studies to quantify the biases of a given source [33, 219]. For cases where content from both biases is shared across the same groups, we label it mainstream. In a preliminary analysis, we found that during the 2018 electoral period right-wing groups were more active in the dissemination of content within WhatsApp. Furthermore, an image posted in those groups is more likely to be associated with a fake story than one posted in another group due to the imbalance between fake and unchecked content. This corroborates with previous studies showing that right-wing groups are more effective in using the social media tool to spread news, disinformation and opinions [35].

4.1.3 Environment Features

Some features can be extracted from the environment such as user engagement metrics and statistics from propagation dynamics.These features have been extensively used in previous efforts [78], especially to better understand the phenomenon of fake news [277]. Next, we detail the features from this category.

Engagement (ENGA). We use measurements from social interactions based on the information available in each of the datasets: (i) in the US election dataset, we use number of likes, shares and comments from Facebook users. (ii) the Health dataset provides information from Twitter users, including number of replies, retweets, etc. Moreover, for both, we compute these numbers within intervals from publication time (900, 1800, 2700, 3600, 7200, 14400, 28800, 57600, 86400, 172800, 259200, 345600, and 432000 seconds), summing up to 12 features. Last, (iii) considering the Brazilian election dataset, we computed measures such as number of distinct users who posted the same news story through image on WhatsApp, the number of distinct groups in which the same news story was posted, and the total number of copies (shares) of the same news story across all analyzed groups both, for messages containing fake news stories and for messages with unchecked content.

External Propagation (EXPR). We also recover external propagation measures, i.e., information about the dissemination of the news stories on the Web. To accomplish this, we use the image associated with piece of news, i.e., the main image in the news story. Next, we use the information about pages with matching images from the Google Vision API which returns information about websites that contain images identical to an image provided as input. From the set of websites/domains that published this image over the Web, we measured the volume of available, uncommon[14], and secure links (i.e., https).

Temporal Engagement Patterns (TEMP). Last, to capture temporal patterns from user activity on digital platforms (i.e., Facebook, Twiiter, and WhatsApp), we compute the rate at which shares/comments are made within intervals for the same time windows defined before.

4.1.4 Novel and Disregarded Features

Despite our efforts to include all the features described before, a few of them could not be included for few reasons. First, some datasets do not contain some pieces of information (e.g., Brazilian election dataset does not contain information related to news URLs – etc). In other words, features were extracted considering the availability of the information in each of the datasets.

More importantly, 22 of the previously described features are novel. In particular, we proposed all features related to domain, including IP, latitude, longitude, and domain credibility. We also proposed other features such as toxicity, safe search indicators, external propagation measure (outside online digital platforms) and readability to assess the writing style of news stories. Later on we show that some of these features were proven valuable for fake news detection.

4.2 Feature Importance

Last, we evaluate the relative power of each feature in distinguishing fake news from others (i.e., true/unchecked content) by ranking them w.r.t. the Information Gain (IG) [21]. Table 4.3 summarizes the results, showing the rank of top-20 most discriminative features according to this measure in each dataset (US and Brazilian elections and Health datasets).

[14] To determine common links we used pre-defined suffixes: '.com', '.net', '.edu', '.org', '.mil', '.gov', '.br' from `https://www.domain.com/blog/2018/10/30/domain-name-types/`

Table 4.3: Feature importance in each dataset.

	Top features by InfoGain (IG)					
	US Election	(%)	Brazilian Election	(%)	Health	(%)
1	ip_to_asn_ASN2635 (source:PUBL)	19.0	count_web_dissemination_urls (env:EXPR)	9.3	toxicity (cont:SEMA)	4.7
2	share_count (env:ENGA)	3.9	web_dissem_accessible_links (env:EXPR)	5.1	count_shares (env:ENGA)	4.0
3	adverb (cont:LEXI)	2.6	web_dissem_foreign_uncom_domain (env:EXPR)	4.3	latitude (source:PUBL)	3.1
4	low_credibility (source:CRED)	1.9	acc_259200 (env:ENGA)	3.8	count_low_word (cont:LEXI)	3.0
5	compare (cont:LEXI)	1.8	count_groups (env:ENGA)	2.9	users_count (env:ENGA)	2.9
6	Dict words from LIWC (cont:LEXI)	1.8	Dict words from LIWC (cont:LEXI)	1.8	relativ (cont:PSYC)	2.3
7	Clout (cont:PSYC)	1.7	Dic (cont:LEXI)	2.1	ranking_position_alexa (source:CRED)	2.2
8	reaction_count (env:ENGA)	1.6	Bridge (cont:SEMA)	2.0	img_count_labels (cont:IMAG)	2.1
9	Period (cont:LEXI)	1.6	ingest (cont:PSYC)	1.9	Dict words from LIWC (cont:LEXI)	2.0
10	senten_info_complex_words_dc (cont:SYNT)	1.5	count_low_word (cont:LEXI)	1.8	sentence_info_characters (cont:SYNT)	2.0
11	readability_index_ARI (cont:SYNT)	1.5	toxicity (cont:SEMA)	1.8	insight (cont:PSYC)	1.8
12	sentence_info_type_token_ratio (cont:SYNT)	1.5	Quote (cont:LEXI)	1.7	seqMatch_top_newsp_alexa (source:CRED)	1.6
13	swear (cont:LEXI)	1.4	img_faces (img_has_faces) (env:IMAG)	1.7	AllPunc (cont:LEXI)	1.5
14	words per sentence (cont:LEXI)	1.3	img_count_labels (cont:IMAG)	1.6	Sixltr, Words > 6 letters (cont:LEXI)	1.5
15	readability_coleman_liau_index (cont:SYNT)	1.3	Sixltr (cont:LEXI)	1.6	web_dissem_accessible_links (env:EXPR)	1.5
16	ppron (cont:LEXI)	1.3	img_count_objects (cont:IMAG)	1.6	readability_dale_chall (cont:SYNT)	1.4
17	verb (cont:LEXI)	1.3	political_bias_right (source:BIAS)	1.6	readability_LIX (cont:SYNT)	1.4
18	average_hashtags_by_comments (cont:LEXI)	1.2	anger (cont:PSYC)	1.5	time (cont:PSYC)	1.3
19	acc_1800 (env:TEMP)	1.1	number (cont:LEXI)	1.4	readability_gunning_fog (cont:SYNT)	1.3
20	ip_to_asn_ASN32244 (source:PUBL)	1.1	cogmech (cont:PSYC)	1.4	function (cont:LEXI)	1.3

Note that the 20 most discriminative features in each dataset are distributed among the three categories, i.e., content, source, and environment, underlining the need to use all of them. Moreover, we observe a trend for all datasets: the content features are the majority, followed by environment and source-related ones. Although the number of features extracted from content is larger, this highlights how important they are for detecting fake news, especially those related to semantic aspects. For instance, there are several news items that are labeled as fake simply because they present information, in some cases even true, but out of context. In summary, the IG results suggesting that our features are useful for fake news detection purposes.

4.3 Summary

In addition to exploring the main features proposed in the literature for fake news detection, in this chapter we presented a new set of features. In total, we computed a total of 199, 177, and 193 features using data from the US election, Brazilian Election and, Health datasets, respectively. We also evaluated the importance of these features for fake news detection using the selected datasets (i.e., data from US and Brazilian elections, and Health) by ranking them w.r.t. the Information Gain (IG). The results reveal that our implemented features can be useful to detect fake news, which we will investigate more deeply and comprehensively in the following sections. Thus, in the next chapter, we use all these implemented features and selected datasets to assess the ability of current supervised machine learning approaches to correctly classify a news story as fake.

Chapter 5

Prediction Performance of Fake News Detection

Although fake news detection is not a new problem[1], recently, this issue gained a lot of strength and its impact in areas, such as politics, has attracted the attention of various researchers who are interested in verifying whether this type of content can, inclusively, manipulate the results of an election [139]. Hence, there is a growing number of studies that attempt to provide a solution to the problem [53, 257, 276, 282].

However, they are mostly concurrent work, which propose complementary solutions and features to train a classifier using data from a specific context, providing hints and insights that are rarely or never tested together. Thus, it is difficult to gauge the potential that supervised models trained from features proposed in recent studies have for detecting fake news. While a fully automated approach for the fake news problem can be quite controversial and is still open for debate, a pertinent research question is: *What is the prediction performance of current approaches and features for automatic detection of fake news in different scenarios?*

To answer this question, in this chapter, we explore the selected datasets from different scenarios (Chapter 3), the main features for fake news detection and the new ones proposed here (Chapter 4) to evaluate and compare different solutions to the problem, assessing the prediction performance of current supervised machine learning approaches. Thus, Section 5.1 describes our experimental setup including metrics used to evaluate our results (Section 5.1.1), the classification methods adopted in this work (Section 5.1.2) and then some other details (Section 5.1.3). Last, the results obtained and a discussion of our findings are presented in Section 5.2. Overall, our results show

[1] `http://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535`

that the prediction performance of features combined with existing classifiers has a useful degree of discriminative power for detecting fake news.

5.1 Experimental Setup

According to the reasons previously mentioned and to our definition of 'Fake News Detection' in Section 2.2.3, in this thesis, we explore a binary classification version task. Thus, the algorithm learns a classification model from a set of previously labeled (i.e., pre-classified) data, and then applies the acquired knowledge to classify new (unseen) news into two classes: fake and non-fake. Based on that, we describe in the next subsections details of the experimental setup adopted in this work.

5.1.1 Evaluation Metrics

To assess the effectiveness of our classification strategy, we adopted metrics commonly used in Machine Learning and Information Retrieval [21, 292]: Precision (P), Recall (R) and F1 by class, MacroF1, and Area Under the Curve (AUC). In this thesis, all these metrics were computed using scikit-learn[2], an open machine learning library in Python. To explain these metrics in our context, we present the following confusion matrix [206], in Table 5.1.

Table 5.1: Example of confusion matrix.

		Predicted	
		Fake	Not Fake/Unchecked
Real	Fake	a	b
Label	Not Fake/Unchecked	c	d

Each letter in the Table 5.1 represents the number of instances that are actually in class X and predicted as class Y, where {X;Y} ∈ {fake and not fake/unchecked}. Recall (R) of a class X is the ratio of the number of elements correctly classified as X to the number of known elements in class X. Precision (P) of a class X is the ratio of the number of elements classified correctly as X to the total predicted as the class X. For example, the precision of the fake news class is computed as: $P(fake) = a/(a+c)$; its recall, as: $R(fake) = a/(a+b)$; and the $F1$ measure is the harmonic mean between both precision and recall. In this case, $F1(fake) = \frac{2P(fake)\cdot R(fake)}{P(fake)+R(fake)}$.

We also compute a variation of F1, namely, Macro-F1, which is normally reported to evaluate classification effectiveness on skewed datasets. Macro-F1 values are

[2] http://scikit-learn.org

computed by first calculating F1 values for each class in isolation, and then averaging over all classes. Macro-F1 considers equally important the effectiveness in *each class*, independently of the relative size of the class. Thus, accuracy and Macro-F1 provide complementary assessments of the classification effectiveness. Macro-F1 is especially important when the class distribution is very skewed, to verify the capability of the method to perform well in the smaller classes, which occurs in our scenario.

In addition, from this confusion matrix in Table 5.1 we can compute the False Positive Rate (FPR) (fall-out) that corresponds to the number of elements in class X that are mistakenly considered as Y, with respect to all elements in class X. Thus, the higher FPR, the more elements in class Y will be missclassified. For instance, FPR of the fake news class is computed as: $FPR(fake) = b/(a+b)$. Then we can combine the FPR and recall (R), also called True Positive Rate (TPR), into one single metric, the Area under ROC curve (AUC). The Area under ROC curve is a metric for binary classification and often used as a measure of quality of the models' performance.

We compute the two first metrics with many different thresholds, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called Area Under Curve (ROC curve), and the metric we consider is the AUC of this curve. The AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen element in class X (positive) higher than a randomly chosen element in class Y (negative).

5.1.1.1 Winning Number

We also resort to a performance measure proposed by Qin Tao et al. [208], called *winning number* to assess the most competitive classifiers (or methods) among candidates, given a pre-defined task they have to perform. In our context, the task is to classify news as fake or not fake/unchecked. That is, the winning number of a method i in the context of a performance measure M, is given as:

$$S_i(M) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbf{1}_{M_i(j) > M_k(j)} \tag{5.1}$$

where j is the index of a dataset, i and k are the indices of a classifier, $M_i(j)$ is the performance (position) of the $i-th$ method on $j-th$ dataset in terms of measure M, and $\mathbf{1}_{M_i(j) > M_k(j)}$ is the indicator function:

$$\mathbf{1}_{M_i(j) > M_k(j)} = \begin{cases} 1 & \text{if } M_i(j) > M_k(j), \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Therefore, the larger $S_i(M)$ is, the better the $i-th$ classifier performs compared to the others. The results of this analysis will be presented in next sections.

5.1.2 Classifiers

We analyzed five different and widely used machine learning algorithms in our experiment: k-Nearest Neighbors (KNN) [57], Naive Bayes (NB) [164], random forests (RF) [31], non-linear Support Vector Machine with the Radial Basis Function (SVM) [122] and, XGBoost (XGB) [46]. We do not include a neural network model for the following reasons: (i) we already have a large number of hand-crafted features for the dataset size – i.e., the datasets that we use are small; (ii) although previous works that employ neural networks exhibited robust performance [283], they lack the explanability achieved with feature engineering that can be useful to understand "the structure of fake news", which we intend to explore in this thesis.

The k-Nearest Neighbors (KNN) is a classification method that performs a selection of k closest training examples in the feature space to assign a label to a unlabeled instance. Naive Bayes (NB) methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. The random forests (RF) classifier is a variation of a bagging of decision trees: it is an ensemble of low-correlated decision trees built by a random attribute selection to compose the decision nodes. The non-linear Support Vector Machine with the Radial Basis Function (SVM) constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification by separating the training instances with maximum distance (margin) in different classes, in our case, fake and not fake/unchecked. Finally, XGBoost (or simply, XGB) is short for eXtreme Gradient Boosting. It is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.

5.1.3 Experimental Details

The predictive performance of models was measured using a 5-fold cross-validation method. The entire 5-fold cross validation was repeated 10 times with different seeds used to shuffle the original dataset, thus producing 50 different results for each test. Thus, the results reported are averages of the 50 runs with 95% confidence intervals.

Last, the best parameters of methods were obtained using a grid search strategy whose implementation, as well as that of all classifiers, is available in scikit-learn[3].

5.2 Classification Results

Table 5.2 shows the empirical results obtained with 95% confidence intervals from the fitted models using all features previously described (see Table 4.2) in each dataset of interest. Overall, we note that the classifiers yield results with some variations across the datasets from different scenarios. Nonetheless, we observe a trend: for all datasets XGB classifier is always among the best performing methods.

Regarding the US election dataset, the best results were obtained by RF and XGB classifiers, statistically tied with 0.80 (±0.015) and 0.79 (±0.019) for MacroF1 and 0.86 (±0.019) and 0.87 (±0.009) for AUC, respectively. Furthermore, XGB is also the best performing strategy considering the Brazilian election dataset, with 0.97 (±0.001) and 0.82 (±0.026) for MacroF1 and AUC, respectively. Last, SVM also appears among the best classifiers in the Health dataset with 0.59 (±0.017) and 0.64 (±0.015) for MacroF1 and AUC. However, it is statiscally tied with XGB classifier that, considering the Health dataset obtained 0.59 (±0.015) and 0.62 (±0.019) for MacroF1 and AUC, respectively.

As we earlier present in Chapter 1 of this thesis, we include the Heatlh dataset as way as to compare our results and measure the potential of our features for fake news detection in a scenario different from politics. However, we note that the proposed and implemented features are not ideal for finding health fake news. The models performed poorly in comparison with results in other datasets, w.r.t. all evaluation metrics, indicating a possible dependency between the task of detecting fake news and the context.

Also, observing F1 by class, we note that all classifiers are better at classifying news as “not fake/unchecked” in comparison with “fake”. This occurred for all datasets analyzed. Note that all datasets provide fewer data labeled as “fake”. We believe that this best represents the real world, where the proportion of fake news fabricated and

[3] `http://scikit-learn.org`

Table 5.2: Experimental results of different classifier algorithms for selected datasets.

Dataset	**Classifier**	**Fake**			**Not Fake/Unchecked**			**MacroF1**	**AUC.**
		P	**R**	**F1**	**P**	**R**	**F1**		
US Election	KNN	0.62±0.187	0.17±0.108	0.23±0.111	0.85±0.012	0.96±0.031	0.90±0.010	0.78±0.014	0.84±0.009
	NB	0.00±0.000	0.00±0.000	0.00±0.000	0.83±0.000	1.00±0.000	0.91±0.000	0.75±0.001	0.81±0.015
	RF	0.51±0.080	0.29±0.138	0.33±0.109	0.86±0.016	0.93±0.045	0.90±0.016	**0.80±0.015**	**0.86±0.019**
	SVM	0.60±0.085	0.12±0.024	0.20±0.036	0.84±0.003	0.98±0.007	0.91±0.003	0.78±0.007	0.84±0.015
	XGB	0.60±0.195	0.15±0.096	0.22±0.115	0.85±0.012	0.97±0.021	0.90±0.004	**0.79±0.019**	**0.87±0.009**
Brazilian Election	KNN	0.00±0.000	0.00±0.000	0.00±0.000	0.97±0.000	1.00±0.000	0.98±0.000	0.95±0.000	0.63±0.030
	NB	0.00±0.000	0.00±0.000	0.00±0.000	0.97±0.000	1.00±0.000	0.98±0.000	0.95±0.000	0.64±0.003
	RF	0.00±0.000	0.00±0.000	0.00±0.000	0.97±0.000	1.00±0.000	0.98±0.000	0.96±0.001	0.71±0.059
	SVM	0.00±0.000	0.00±0.000	0.00±0.000	0.97±0.000	1.00±0.000	0.98±0.000	0.95±0.001	0.46±0.105
	XGB	0.40±0.043	0.07±0.017	0.03±0.030	0.97±0.000	1.00±0.000	0.98±0.004	**0.97±0.001**	**0.82±0.026**
Health	KNN	0.57±0.132	0.04±0.011	0.07±0.019	0.67±0.003	0.98±0.012	0.80±0.004	0.56±0.006	0.61±0.013
	NB	0.00±0.000	0.00±0.000	0.00±0.000	0.67±0.000	1.00±0.000	0.80±0.000	0.53±0.001	0.55±0.024
	RF	0.35±0.164	0.07±0.080	0.11±0.010	0.68±0.011	0.96±0.044	0.79±0.010	0.56±0.028	0.61±0.019
	SVM	0.56±0.055	0.11±0.031	0.18±0.048	0.68±0.008	0.96±0.011	0.80±0.006	**0.59±0.017**	**0.64±0.015**
	XGB	0.50±0.059	0.12±0.059	0.19±0.070	0.68±0.005	0.93±0.042	0.78±0.011	**0.59±0.015**	**0.62±0.019**

shared is small compared to the large volume of news generated every day. However, with less data labeled as "fake" to train the classifiers, they are naturally worse at distinguishing them from others, highlighting that detecting fake news is not a trivial task.

Last, as an initial attempt to investigate the practical potential of these approaches for detecting fake news, we inspect the ROC curve for the best classifiers in each dataset as shown in Figure 5.1. We observe that it is possible to choose a threshold to correctly classify most of fake news (true positive rate ≈ 1), while misclassifying 40% and 55% of the true/unchecked content considering data from the US and Brazilian elections, respectively (false positive rate ≈ 0.40 and ≈ 0.55). For the Health dataset, the results reinforce that the models are not useful for correctly identifying health fake news.

In the next section, we present the results of the winning number score achieved for all classifiers in the labeled datasets.

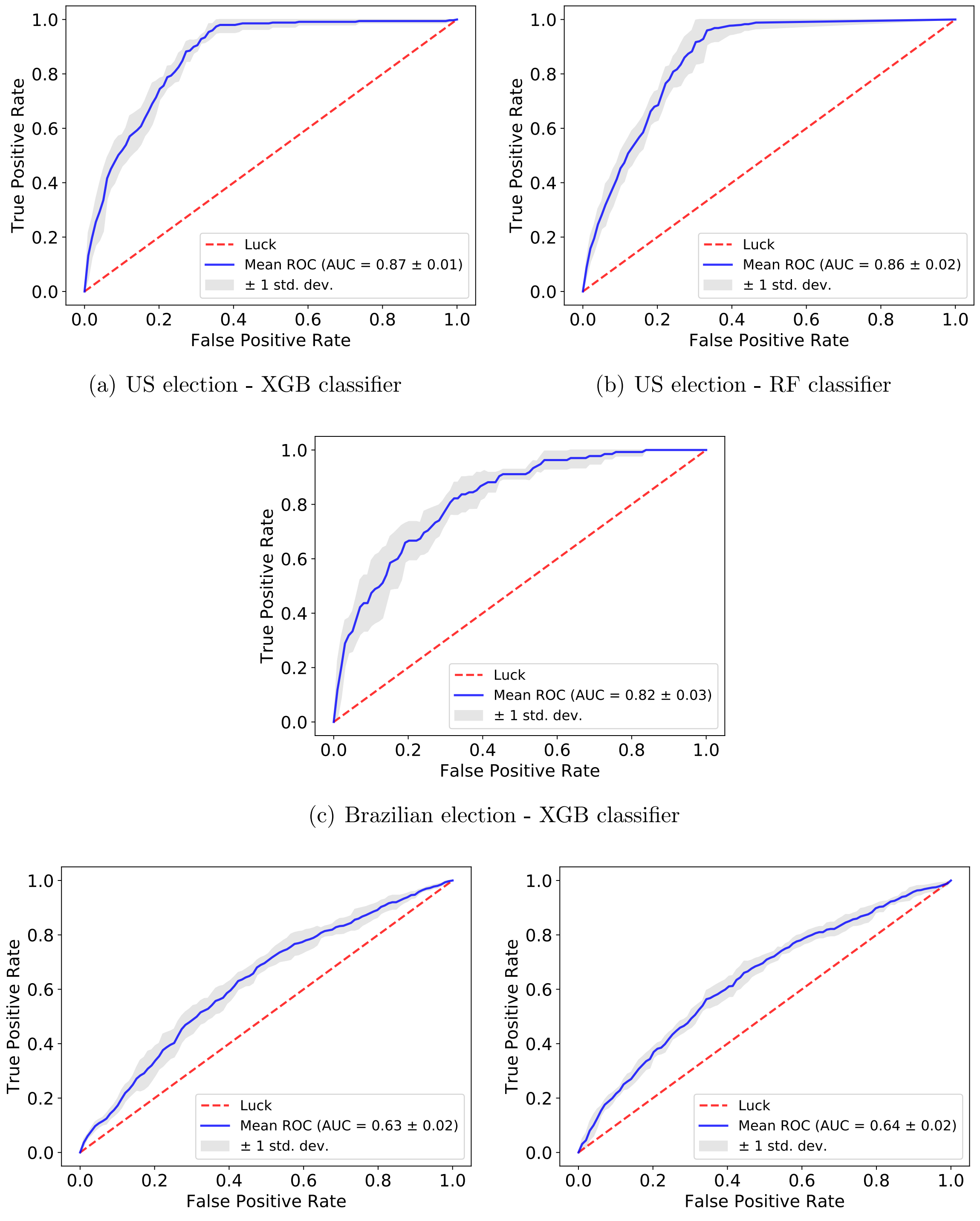
5.2.1 Winning Numbers

The winning number measure tries to assess the most competitive methods among a series of candidates, given a large series of pre-defined tasks they have to perform. By Equation 5.1, the highest winning number that could be achieved by each classifier is 15. Table 5.3 presents the results of winning score, in which we consider the performance metric MacroF1 and AUC.

In Table 5.3, the top classifier is XGB, followed by RF, SVM, KNN, and NB. This means that XGB method performs well across datasets when it comes to correctly distinguish fake news from others (i.e., not fake/unchecked). This suggests that XGB would be preferable in situations in which a preliminary evaluation has to perform. However, it is important to note that some supervised classification results are considerably low (e.g., considering the Health dataset), still leaving a large gap for the development of better solutions for fake news detection.

Table 5.3: Winning points ranking for MacroF1 and AUC.

Classifier	MacroF1 Winning Score	AUC Winning Score
XGB	15	15
RF	12	12
SVM	11	9
KNN	9	8
NB	5	5



(a) US election - XGB classifier

(b) US election - RF classifier

(c) Brazilian election - XGB classifier

(d) Health - XGB classifier

(e) Health - SVM classifier

Figure 5.1: For the US and Brazilian elections datasets, it is possible to correctly classify almost all of fake news with only 40% and 55% of false positive rate, respectively. Regarding Health dataset, the results show that the models are not useful for correctly identifying health fake news.

5.3 Summary

In this chapter, we explored the prediction performance of current approaches and features for the automatic detection of fake news. Particularly, considering our surveys on existent datasets and features for fake news detection in Chapter 4, we explored the selected datasets, including the one proposed here, the main features for fake news detection, and proposed new ones to evaluate and compare different supervised machine learning approaches, assessing their prediction performance in the task of automatically identify fake news in different scenarios. Our results provide an interesting perspective on the current prediction performance of supervised machine learning approaches for fake news detection. For instance, our best classification results for the US election dataset considering MacroF1 (i.e., RF and XGB, statistically tied with 0.80 ± 0.015 and 0.79 ± 0.019, respectively and AUC (i.e., 0.86 ± 0.019 and 0.87 ± 0.011 for RF and XGB, respectively) as evaluation metrics reveal that these classifiers are accurate for the fake news detection task. The same can be observed for the Brazilian election dataset (i.e., 0.97 ± 0.001 and 0.82 ± 0.026 obtained by XGB, considering MacroF1 and AUC respectively). Moreover, inspecting the ROC curve, we note that we are able to correctly classify nearly all of the fake news in the US and Brazilian elections datasets, misclassifying about 40% and 55% of true/unchecked news stories, which is already sufficient to help fact-checkers, especially in the identification of news stories that are worth investigating, reducing their search space for suspicious news. Last, our results also reveal that the implemented features are not useful to identify health fake news. Hence, we did not include the Health dataset in the rest of the analyses of this thesis.

Overall, although the prediction performance of proposed features combined with existing classifiers already has a useful degree of discriminative power for detecting fake news from political context, there is a still large space for improvement. First, little is known about the discriminating power of features for fake news detection, either individually or when combined with others. Moreover, explaining the predictions made by classifiers can be useful to the understanding of the influence of each of the features on the outcomes of the classifiers helping the experts in the fact-checking process. We explore these explanations in the next chapter. Particularly, in order to provide a better understanding of the fake news phenomena, we explore the XGB classifier that obtained the best results among a set of classifiers tested and propose a framework for quantifying the informativeness of features for fake news detection using the US and Brazilian elections datasets. In addition, we provide explanations on how these features are used in the decisions taken by computation models designed to detect fake news, allowing us to understand their utility considering different scenarios in the political context.

Chapter 6

Informativeness of Features for Fake News Detection

As previously introduced in this thesis, in our study on RG1, we conducted a systematic survey that identified existing features for fake news detection (see Chapter 4). In addition, we proposed new ones and explored the prediction performance of current supervised machine learning approaches using all these features combined considering data from different scenarios. Our results reveal that the use of supervised machine learning for this task is promising, and the proposed solutions can be used especially in assisting fact-checkers in identifying stories that are worth investigating.

However, little is known about the discriminative power of features proposed in the literature, either individually or when combined with others. Some may be adequate for pinpointing fake news with specific patterns, while others are more general but not discriminative. Moreover, while explaining the decisions made by the proposed models is central to understanding the structure of fake content, this discussion is often left aside. In this chapter, we address all these issues.

Specifically, towards our second research goal (see RG2 in Chapter 1), we want to provide answers to the following questions: *Do we need all features for fake news detection, or should we focus on a smaller set of more representative features? Is there a trade-off between feature discriminative power and robustness? Is there a clear link between features and the patterns of fake news they can detect?*

To answer these questions, we propose a framework for quantifying the informativeness of our implemented features for fake news detection (see Chapter 4) considering our data from the US and Brazilian elections. Since these features may have a variety of complex nonlinear interactions, we employ a fast and effective classification algorithm and with significant flexibility. Finally, we propose and perform an unbiased search for models, so that each model is composed of a set of randomly chosen features. We enumerated roughly 400K and 300K models for each scenario analyzed (i.e., the US and Brazilian elections, respectively), enabling to perform a unique macro-to-micro investigation of the considered features.

Based on our proposed framework for quantifying the informativeness of features, our analysis unveils the real impact of a slew of features for fake news detection considering the scenarios analyzed. Particularly, our results show that: (*i*) our unbiased model exploration reveals how hard fake news detection is, as regarding the 2016 US election dataset, only 2.2% of the models achieve a detection performance higher than 0.85 in terms of the area under the ROC curve (AUC). On the other hand, using data from the 2018 Brazilian election, the models achieved detection performance lower than 0.85 (the best model achieved 0.84 in terms of AUC); (*ii*) among the best models, we found that some features appear up to five times more often than others; (*iii*) we distinguish a small set of features that are not only highly effective to build high-performance models, but also contribute the most to increase the robustness of the models in the different scenarios, and; (*iv*) we represent models in a high dimensional space, so that models that output similar decisions are placed close to each other. We then cluster the model space, and a centroid analysis reveals that prototype models are very distinct from each other. Our cluster analysis by AUC corroborates with these results. For centroids prototypes, we present an explanation of factors contributing to their decisions. Our findings suggest that models within different groups separate fake from real content based on several different reasons.

Additionally, we emphasize that this study is not about proposing the best combination of features or the best classification model, but rather about investigating features' informativeness and simple models that can be generated from them to identify fake news in different scenarios, as well as using these models to explain predictions made for news stories.

Next, we describe our framework for quantifying the informativeness of features for fake news detection, including experimental setup. Finally, we present our results and implications of our findings.

6.1 Framework for Quantifying the Informativeness of Features for Fake News Detection

Figure 6.1 presents an overview of our framework for quantifying the informativeness of features for fake news detection. The framework can be divided into three main

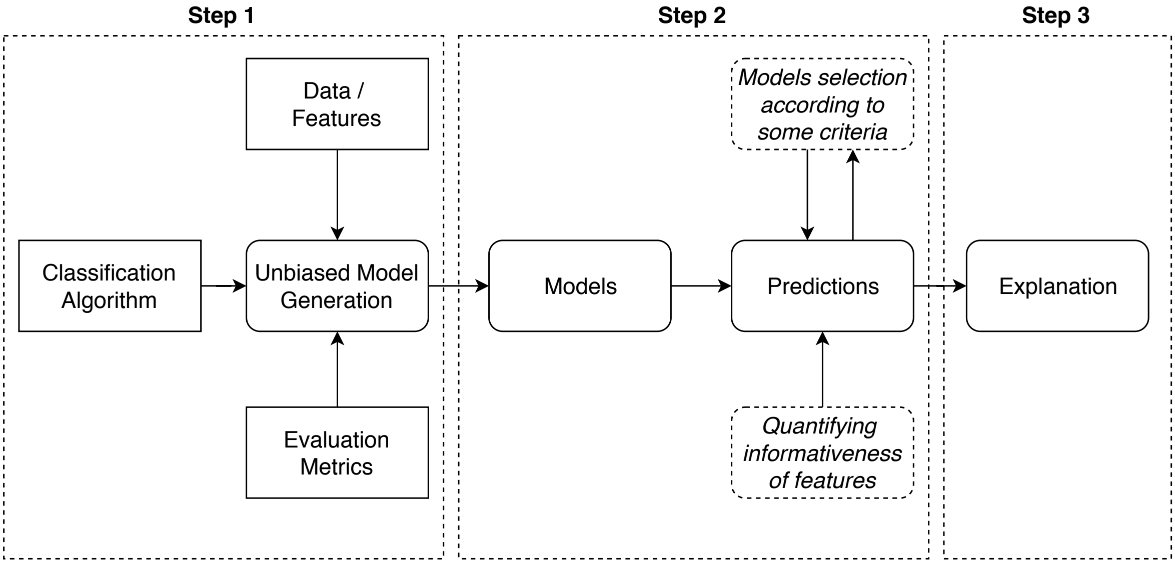

Step 1
Step 2
Step 3
Data / Features
Classification Algorithm
Unbiased Model Generation
Evaluation Metrics
Models
Predictions
Models selection according to some criteria
Quantifying informativeness of features
Explanation

Figure 6.1: Overview of our framework for quantifying the informativeness of features for fake news detection.

steps: **Step 1** involves unbiased generation of models from some inputs: a classification algorithm, the implementation of features using a specific dataset and the definition of metrics that will be used to evaluate models' performance.

Then, **Step 2** consists of quantifying informativeness of features based on models' predictions. Additionally, this step can be executed considering some additional filter or criteria for the selection of models that will be analyzed. For instance, the model with the best performance according to some evaluation metric or representative models of clusters.

Finally, **Step 3** provides an explanation of factors contributing to model decisions, thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions. Next, we describe how we tackle each of these steps in this thesis.

6.2 Step 1: Unbiased Model Generation

There are some inputs for unbiased model generation, as introduced previously by Figure 6.1, that we hereafter present.

Features and data: We consider our 199 and 177 implemented features for fake news detection (see Chapter 4) using the US and Brazilian elections datasets, respectively.

Classification algorithm: The features we consider may have a variety of complex nonlinear interactions. Capturing these interactions requires a classification algorithm with significant flexibility. For this reason, we chose gradient boosting machines. The main idea of gradient boosting machines is to combine multiple models into a stronger one. Specifically, models are iteratively trained so that each model is trained on the errors of the previous models, thus giving more importance to the difficult cases. At each iteration, the errors are computed and a model is fitted to these errors. Finally, the contribution of each base model to the final one is found by minimizing the overall error of the final model. Fitting the base models is computationally challenging so we used a recent, high-performance implementation of gradient boosting machines, the XGB [46], which was the best performing method according to our experiments in the last chapter.

Evaluation: In order to evaluate how accurate the learned models are, in this section we compute the area under the ROC curve (or simply, AUC [21]), which considers the precision-recall trade-off. As mentioned in Section 5.1.1, the AUC is an estimate of the probability that a model will rank a randomly chosen fake news case higher than a randomly chosen fact case. The AUC is robust to class imbalance and considers all possible classification thresholds.

Our unbiased model generation: The exact approach to assess the real impact of features for fake news detection would require exhaustive enumeration of all possible combination of features, so that one model is obtained for each combination of features. Obviously, inspecting all possible subsets of features is computationally prohibitive. Instead, we sample the model space by randomly selecting the features that compose a model considering each of the analyzed dataset.

Then, according to Figure 6.1, in **Step 1** we propose a framework to explore the space $\mathcal{M}$ of models that can be generated from a set of features considering a specific dataset. Even if we consider a single class of models, fix the model parameters, and limit the maximum number F of features in a model, the set of possible models is still $2^F - 1$. Strategies based on forward or backward feature selection favor features selected early on and thus are inadequate for assessing the relative importance of features in the model space. Sampling models uniformly from this space is not desirable either, as models with different number of features may have very different sampling probabilities. Our framework compensates for the lower prevalence of very small and very large models by sampling the same number of models for different number of features. Moreover, each feature occurs exactly the same number of times in the generated model set.

Specifically, we begin by enumerating all possible 1-feature and 2-feature models

(199 and 19,701 models in the US election dataset, and, 177 and 15,576 models in the Brazilian election dataset, respectively). Next, we take each of the 2-feature models and include one new feature chosen uniformly at random, so as to build 3-feature models. This step is repeated until we reach models composed of 20 features[1] (a total of 374,518 and 296,121 models considering the US and Brazilian elections datasets, respectively). In each step, we ensure that each feature is included the same number of times and that no feature appears twice in a model. This compensates for the smaller number of few-feature models by keeping the number of models constant regardless of the number of features. We further present that this number of features (i.e., up to 20) allows us to generate models with better performance compared to the best results we obtained in the previous chapter.

Experimental setup: Similar to our experiments presented in the last chapter, for each XGB model, we perform a 5-fold cross-validation. The dataset is partitioned into five partitions, out of which four are used as training data, and the remaining one is used as the validation-set. The process is then repeated five times with each of the sets used exactly once as the validation-set, thus producing five results. Hence, the reported AUC values are averaged over the five runs. Furthermore, we employ the mean absolute deviation (MAD) in order to get sense of how spread out the AUC values are through the five validation sets. Therefore, for each XGB model, we have an estimate of its predictive accuracy and variability.

6.3 Step 2: Informativeness of Features for Fake News Detection

In this section, we describe the results of the experiments as part of the **Step 2** of our proposed framework for quantifying the informativeness of features for fake news detection using data from the 2016 US and 2018 Brazilian elections. They are designed to answer our research questions. In Section 6.3.1, we investigate the predictive accuracy and variability of features. Then, in Section 6.3.2, we focus on the best performing models in order to evaluate models in terms of effectiveness and variability. Finally, in Section 6.3.3, we cluster the model space according to the features in each model, and we conduct an investigation to understand the role features play in the model decisions.

[1] This number was defined based on the experimental analyzes available in the Appendix B of this thesis.

6.3.1 Features: Accuracy and Variability

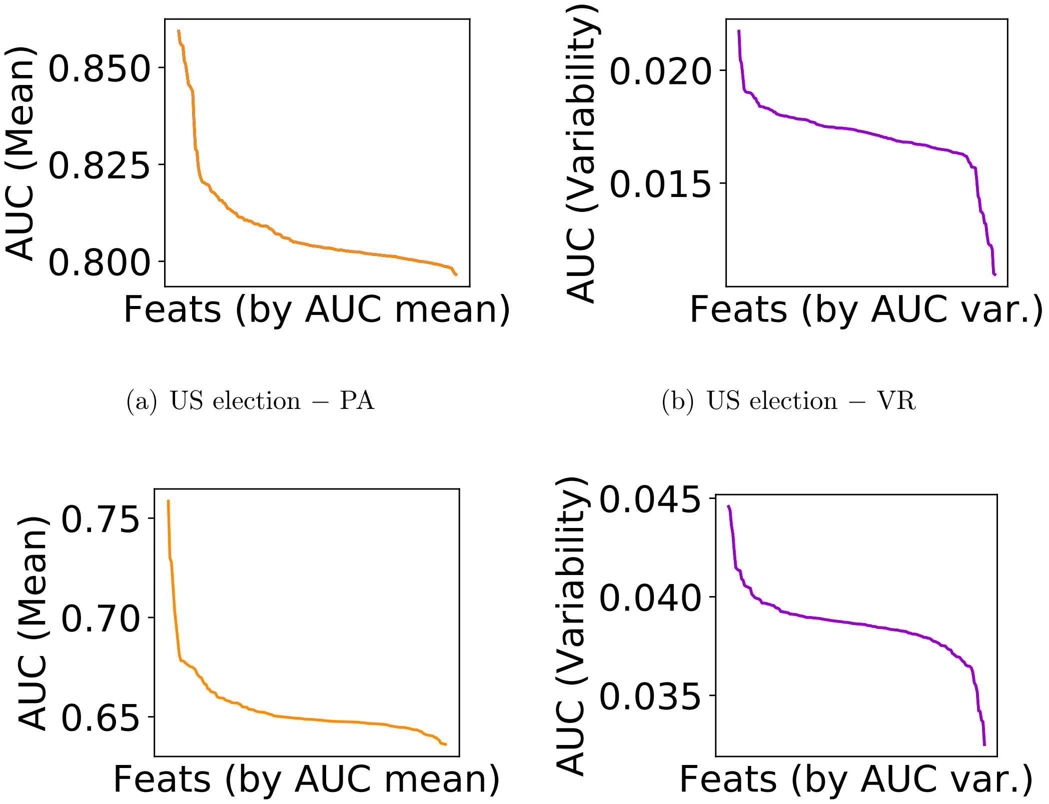
We quantify the predictive accuracy of a feature by considering all models in which the feature was included. Particularly, the predictive accuracy of a feature is given as the average AUC value of all models in which the feature was included. Similarly, the variability of a feature is given as the average MAD value of all models in which the feature was included. Figure 6.2 shows how features are distributed in terms of predictive accuracy and variability considering the US and Brazilian elections datasets. We observe that there is a small number of features for which the predictive accuracy is significantly higher. Specifically for the US election dataset (Figure 6.2(a)), around 3% of the considered features are included into models in which the average AUC values are higher than 0.85. For the Brazilian election dataset (Figure 6.2(c)), approximately 3% of these features are included into models with AUC $>$ 0.70, on average. In both datasets, most features are associated with significantly lower average AUC values. The same trend is observed when we investigate the distribution of features in terms of variability. Around 2% of the considered features are associated with relatively low variability in the US election dataset (Figure 6.2(b)). Considering the Brazilian election dataset, as shown in Figure 6.2(d), this percentage is slightly higher (around 3%).

6.3.2 Top 10 % Models: Accuracy and Variability

Now we investigate whether relatively simple models (composed by up to 20 features) can perform consistently well across the datasets. In order to do so, we take the top 10% models w.r.t. AUC. Among the best performing models, we are interested in those that exhibit low variability.

Figure 6.3 shows a scatter plot of the top 10% models w.r.t AUC, each represented by a dot. Each dot's diameter is proportional to the ratio between the respective model's AUC mean and variability. We show 2D t-SNE representations [160] for the sake of easy visualization. Cartesian coordinates of each dot center are obtained from the vector of the probabilities assigned by the model to each fake news case in the validation set.

First, we note that the mean AUC is in the range $[0.858, 0.885]$ and $[0.738, 0.839]$ for the US and Brazilian elections datasets, respectively. Therefore, the different diameters are mostly due to AUC's variability. For the US election dataset, we observe the presence of very few models with excellent performance on average (yellow dots, AUC $>$ 0.88), but with high variability. On the other hand, there are many models with lower variability, but with lower average AUC values (medium-sized purple and blue



(a) US election – PA

(b) US election – VR

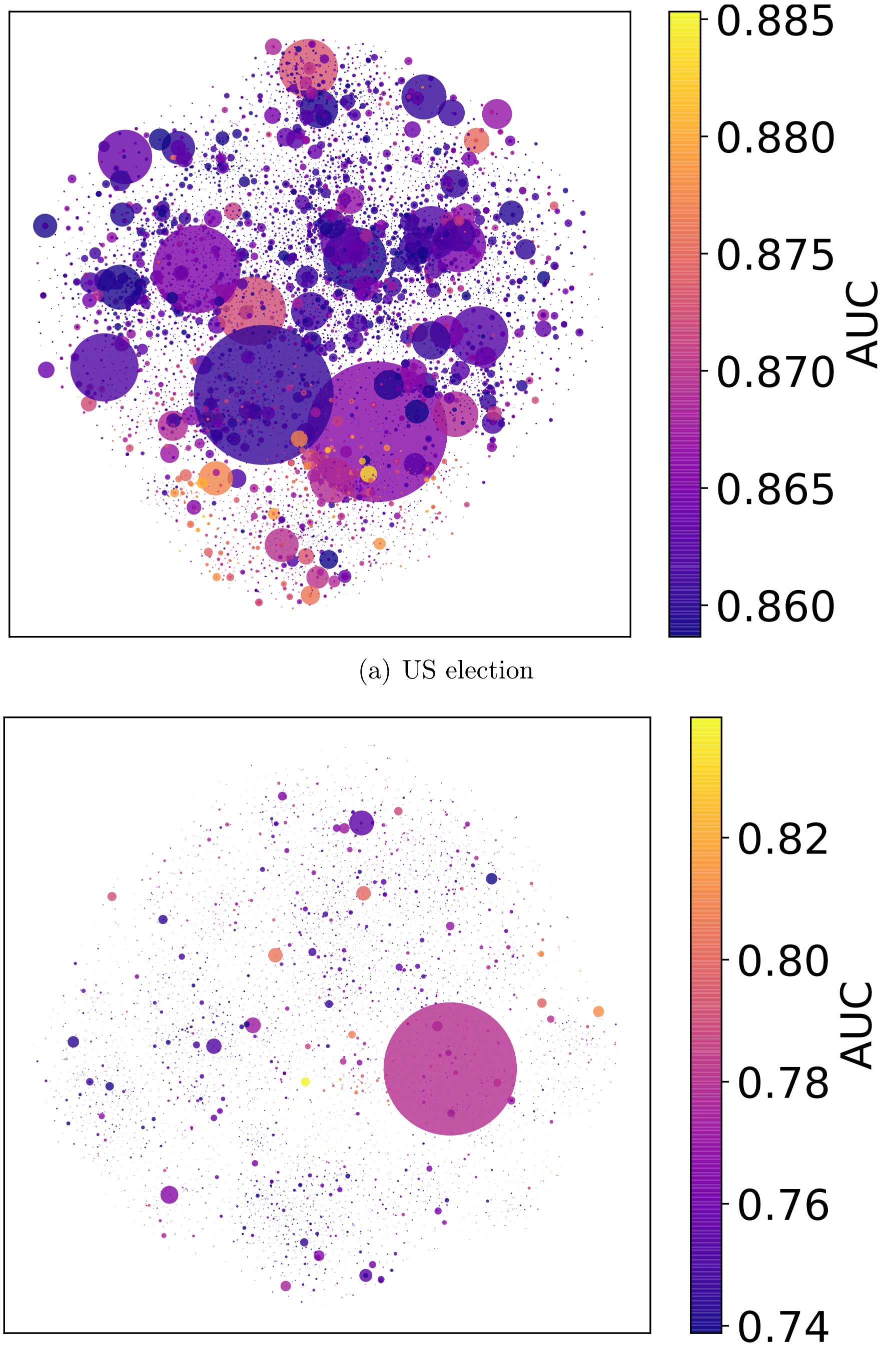
(c) Brazilian election – PA

(d) Brazilian election – VR

Figure 6.2: Distribution of features. Left – Predictive accuracy (PA). Right – Variability (VR).

models, AUC < 0.865). Finally, there are some models with a good trade-off between performance and variability (pinkish dots, AUC ≈ 0.87). Considering the Brazilian election dataset, we observe the presence of few models with good performance on average (yellow dot, AUC > 0.83 and orange dots, AUC > 0.80). Also, there are many models with lower average AUC values (blue and purple dots, AUC < 0.775). Overall, the models have high variability. There is a single model with lower variability but also with lower average AUC value (pinkish dot, AUC ≈ 0.79). We further discuss these cases.

To better understand the relationship between features and model performance in each of the datasets, we take the best performing models and compute the prevalence of features. Also, to understand the relationship between them and AUC variability, we take, from the top performing models, the 10% with the highest and 10% with the lowest variability and compute the prevalence of features in these sets. We present the results in the next sections.



(a) US election

(b) Brazilian election

Figure 6.3: Each point corresponds to a model. Color indicates the AUC value. The diameter indicates variability (i.e., the larger the diameter the lower is the variability). Models are placed according to the probabilities they assign to fake news cases. We show 2D t-SNE representations for the sake of easy visualization.

6.3.2.1 Accuracy

Considering models with the highest AUC values (the top 10% best models) and the following analyzed datasets, we discover the following.

US election dataset. Features that capture information regarding news publisher, including location and credibility of domains (e.g., publisher/domain (54%), ranking position of domain from Social Alexa (39%), page reactions from users on digital platforms (22%), IP (ASN) associated with domain (20%), etc.), are very frequent in this group of models. Moreover, features extracted from the environment (i.e., from digital platforms), are also more frequent (e.g., the number of shares (35%) and reactions count (29%)). Finally, features that capture political bias of news outlet (i.e., mainstream, left-leaning, and right-leaning (39%)) are quite prevalent in the best models. On the other hand, character level, word-level and sentence-level features (e.g., count) are less frequent in best models (7% of models on average).

Brazilian election dataset. Features extracted from the environment (i.e., from digital platforms and Web) are more frequent (e.g., information about the dissemination of the stories on the Web (80%), number of Websites that published the news story over the Web (38%), number of shares (17%) and users that disseminated the news story (20%), etc). Also, information regarding semantic structure of news story (e.g., image labels (22%), toxicity (18%), image safe search (17%), etc) and news publisher (e.g., the first group in which the news story was posted (17%)) are very frequent. Last, character level, word-level and sentence-level features (e.g., number of articles, prepositions and punctuations) are less frequent in best models in the Brazilian election dataset (8% of models on average).

6.3.2.2 Variability

In terms of variability among the best models, we observe that:

US election dataset. While features from the digital platforms (e.g., share count (8%)) and readability indicators (8%) are very frequent in models with low variability, features from user engagement (from digital platforms) and source (e.g., Facebook page (5%) and political bias (6%)), occurred more often in models with high variability. Word level-features are very frequent both in models with high and low variability.

Brazilian Election Dataset. Features that capture information regarding subjectivity of text (8%), semantic structure (8%), political bias (7%) and information about the dissemination of the news stories on the Web (8%) are very frequent in models with low variability. On the other hand, features extracted from the environment (e.g., number of shares on first 432000 seconds (6%)), are frequent in models with high variability. Also, Word level-features and sentence-level features (e.g., readability indicators) are very frequent both in models with high and low variability.

In summary, we conclude that there are many combinations of features that yield models with high performance and low variability. In the next section, we investigate whether these models are redundant (i.e., identify similar sets of fake news) or complementary.

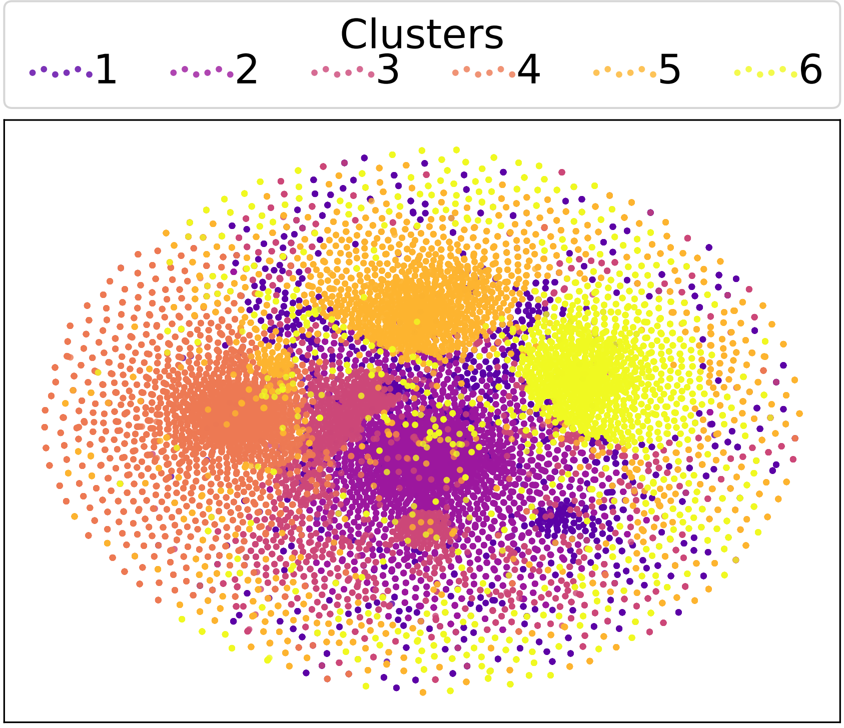
6.3.3 Clustering the Model Space

In order to understand whether the top 10% models cover different regions of the space of fake news considering the two datasets analyzed, we cluster them from one hot vector representations that indicate which features are present in each model. To cluster these models, we use the standard K-Means algorithm based on Euclidean distances [16], which divides the set of coalitions into K disjoint clusters. The distance between the coalitions is given by their euclidean distances. To find the optimum value of K, we use the Silhouette Score [226], which measures, on average, how tightly grouped all the members in different clusters are, and selects the value of K, for which the Silhouette Score is the highest. In this thesis, we use $K = 6$ for the US election dataset, and $K = 4$ for the Brazilian election dataset. The sizes of the resulting clusters vary from 4,012 to 8,391 (mean 6,241 and std. dev. 1,529) and from 2,908 to 8,517 (mean 5,922 and std. dev. 1,846) for the US and Brazilian elections datasets, respectively.

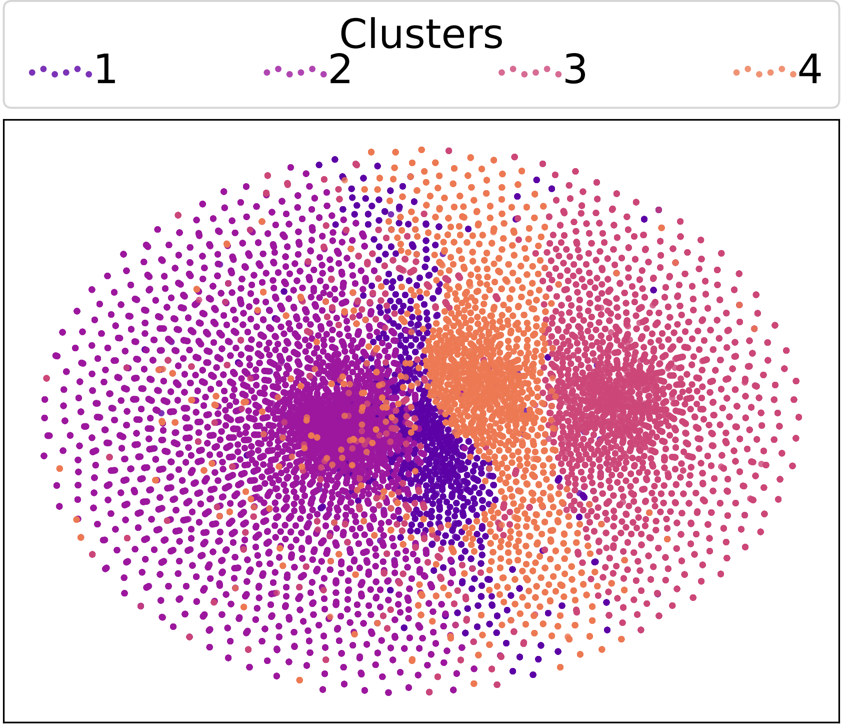
Once again, we embed the models in a 2D space based on the probabilities assigned to fake news cases and color code the models according to the cluster they belong to. Hence cohesive clusters indicate that models within the same cluster are better at identifying fake news with specific patterns. If this is the case, models that belong to the same cluster (i.e., share similar features) are expected to be close to each other in the embedding, indicating that they assign similar scores to the fake news in the test set. In fact, this is what we observe in Figure 6.4 for both datasets. Next, we analyze which types of features best describe each cluster.

6.3.3.1 Describing Clusters in Terms of Types of Features

When we focus on the analysis of the top 10% performing models in each dataset, features no longer appear with the same frequency. In addition, clusters include different

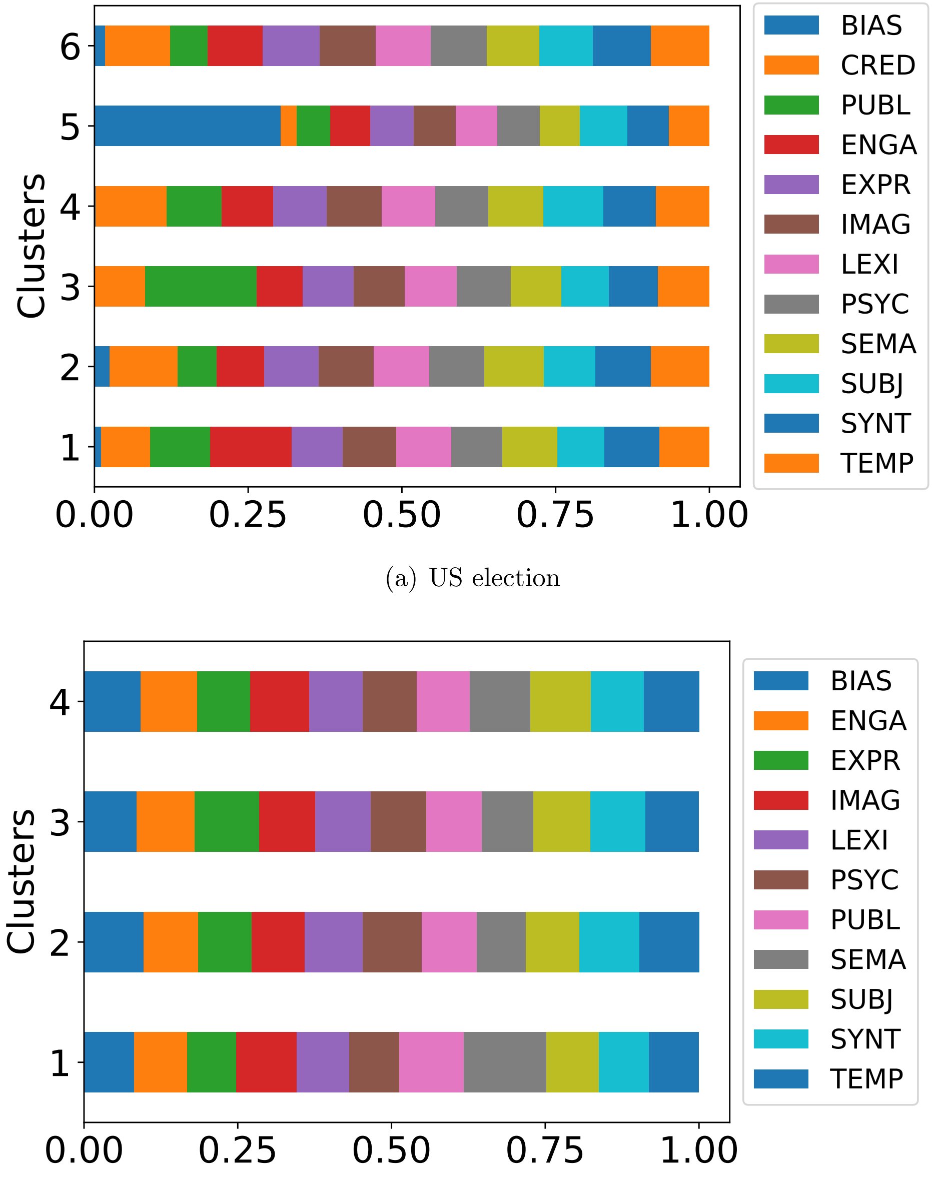


(a) US election



(b) Brazilian election

Figure 6.4: t-SNE representations of models based on the scores associated to each fake news in the validation set in each dataset. Colors indicate clusters found from binary vectors indicating which features were used in each model. Proximity between models from the same cluster suggest correlation between features used and fake news correctly detected.



(a) US election

(b) Brazilian election

Figure 6.5: Descriptions of clusters in terms of feature sets, represented as segments. Segment lengths are normalized $R_{i,t}$ ratios and indicate how much more/less often features of type t appear in cluster i than in a random null model.

number of models, each of which can include any number of features. In order to compare the frequency of specific types of features across clusters, we define a (random) null model. This allows to determine how much more (less) often than expected a given feature type appears in a cluster.

Let $C_{i,t}$ be the number of times features of type t appear in models from cluster i for $t \in$ BIAS, CRED, ..., TEMP and $i = 1, \dots, 4$ or 6 (considering the Brazilian and US elections datasets, respectively). Multiple features of the same type are counted multiple times. Also, let $C_i = \sum_t C_{i,t}$ be the total number of features in cluster i. Denote by $N_t = \sum_i C_{i,t}$ the number of times features of type t appear among the top 10% performing models. The expectation of $C_{i,t}$ if features were assigned to clusters completely at random is $C_i N_t / \sum_t N_t$. Therefore, the ratio $R_{i,t} = C_{i,t}/(C_i N_t / \sum_t N_t)$ measures how much more (less) often features of type t appear in cluster i than in a random null model.

Figure 6.5 shows ratios $R_{i,t}$ normalized for each cluster i (i.e., divided by $\sum_t R_{i,t}$). Normalized ratios allow to identify which types of features better describe each cluster considering both datasets. Regarding the US Election dataset (Figure 6.5(a)), we note that clusters comprise combinations of features types in different proportions. Also, we observe that all clusters use features from all features types, except for clusters 3 and 4, which do not include BIAS features. These features are more frequent in cluster 5 and less so in clusters 2, 4 and 6. CRED features are very prevalent in clusters 2, 4 and 6, but less used by models in cluster 5. Finally, PUBL features are very prevalent in cluster 3. In contrast, for the Brazilian Election dataset (Figure 6.5(b)), we observe that all clusters use features from all features types but in similar proportions, except for cluster 1, which uses SEMA features in a higher proportion than the other clusters. Therefore, especially for the US Election dataset, these observations combined with Figure 6.4 corroborate the hypothesis that models generated from different combinations of features are able to correctly identify different fake news groups.

6.4 Step 3: Feature Importance and Shapley Additive Explanations

Effective models perform decisions that are usually hard to explain. However, understanding why a model has made a specific decision is paramount in any fake news detection application scenario, as it provides insights into the reasons why the content was considered to be fake, tooling fact-checkers with the aspects that contributed most to the decision.

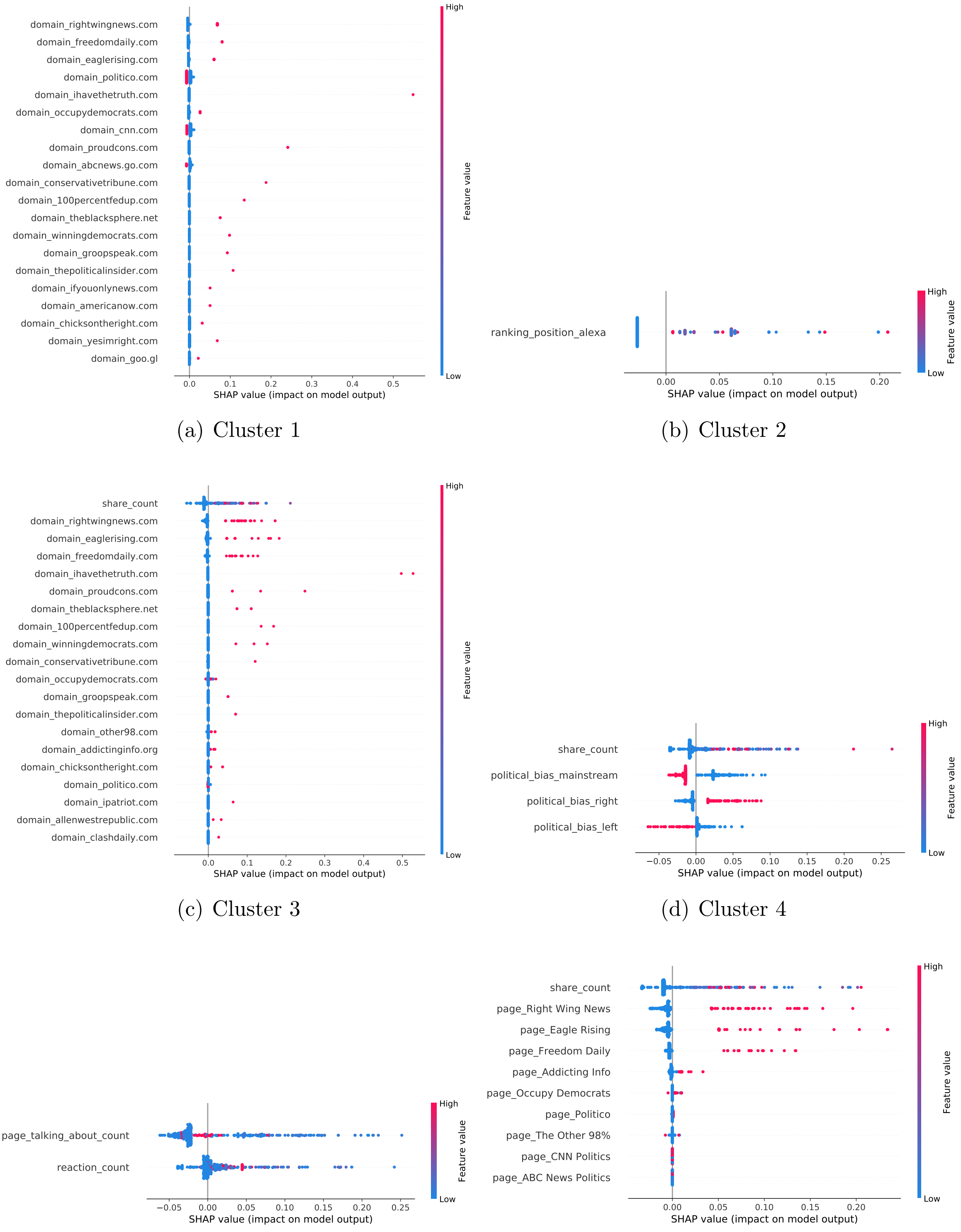
The typical approach for explaining the decisions of a model is based on calculating the impact (or importance) each feature has on the decision. Feature importance can be defined as the increase in the model prediction error after feature values are permuted, since this operation breaks the relationship between the feature and the outcome. Therefore, a feature is important if permuting its values increases the model error, because the model relied on the feature for the decision. On the other hand, a feature is not important if permuting its values keeps the model error unchanged, because the model ignored the feature for the decision.

However, features often interact with each other in many different and complex ways in order to perform accurate decisions. Thus, the feature importance is also given as a function of the interplay between the features. In this case, Shapley values [157] can be used to find a fair division scheme that defines how the total importance should be distributed among the features. In fact, Shapley values are theoretically optimal and the unique consistent and locally accurate attribution values. Unfortunately, Shapley values can be challenging to compute, and thus we focus on explaining only the top-most effective models. Therefore, we attempt to explain the decisions made by some prototype models in Section 6.4.1.

6.4.1 Explaining Model Decisions

In this section, we use SHAP [157] to explain why news are classified as fake or not by representative models of each cluster considering the two datasets analyzed. SHAP is short for SHapley Additive exPlanations. It is a unified approach to interpreting model predictions. As such, SHAP assigns a "force" or importance value – positive or negative – to each feature in a particular prediction [157]. The output value (prediction) consists of the sum of the base value (average prediction over the validation set) and these forces (closer to 1.0 means more likely to be fake). In addition, SHAP allows (i) to summarize the importance of a feature, and (ii) to associate low/high feature values to an increase/decrease in output values, through color-coded violin plots built from all predictions.

Representative models of each cluster in each of the datasets were selected according to the following criteria: (i) by centroid proximity, selecting the closest model to the cluster centroid (Figures 6.6 and 6.7, for the US and Brazilian elections datasets, respectively); and (ii) by AUC, selecting each cluster the model with the best performance w.r.t AUC (Figures 6.8 for the US election dataset and 6.9 for the Brazilian election dataset). In this section, we show only the graph for the first round of cross-validation, in Appendix C, all the graphs for all folds for both criteria, i.e., (i) and (ii),



(a) Cluster 1

(b) Cluster 2

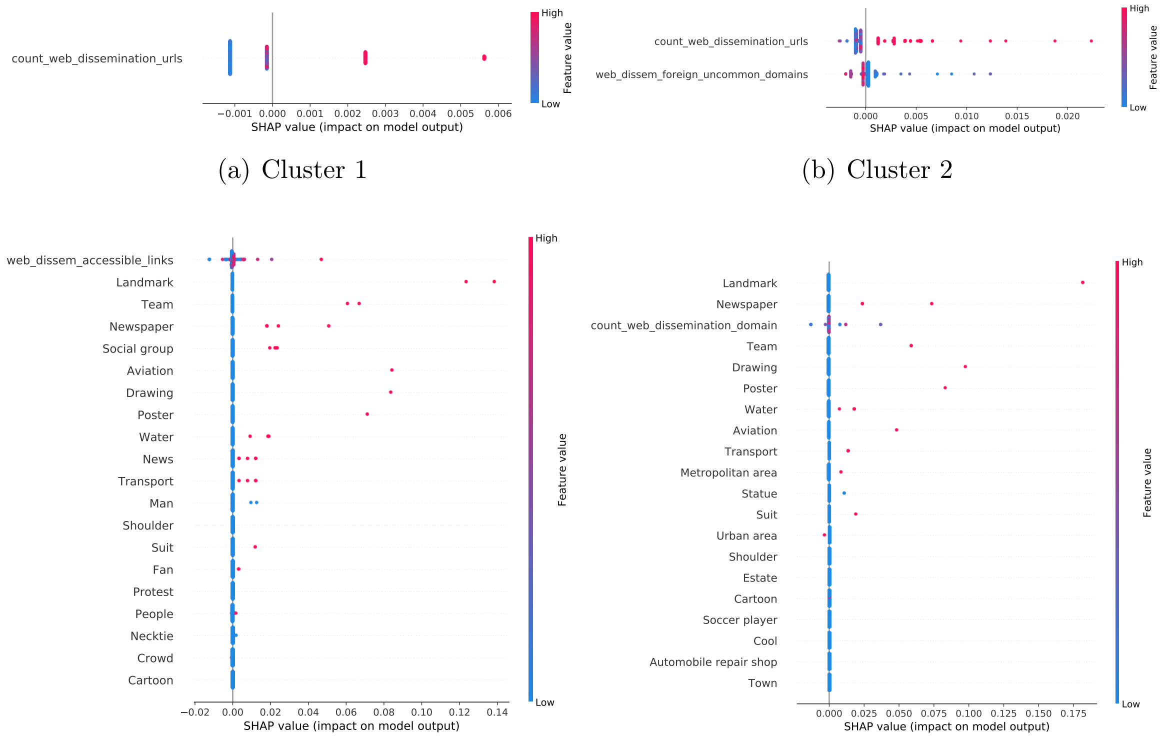
(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Figure 6.6: SHAP summaries for the closest models to each cluster centroid for the US election dataset.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

Figure 6.7: SHAP summaries for the closest models to each cluster centroid for the Brazilian election dataset.

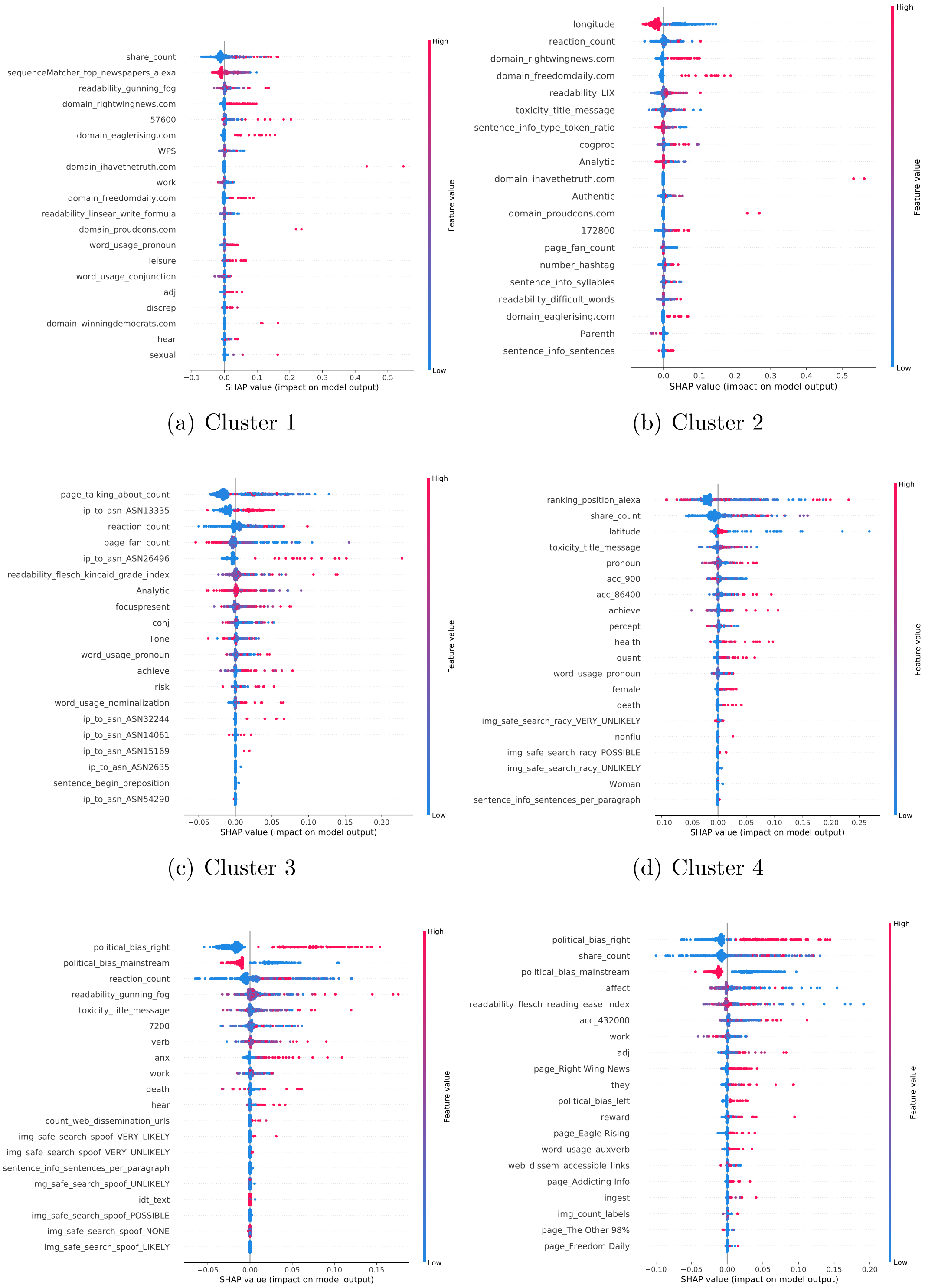
are shown.

Figures 6.6 and 6.7 show violin plots of SHAP values for features used by each of the selected centroid models for the US and Brazilian elections datasets, respectively. Interestingly, we note that the closest models to the centroids for both datasets have either one or two features. Also, for the US election dataset all features of the selected centroid models come from sets PUBL, BIAS, ENGA and CRED. In contrast, for the Brazilian election dataset, these features come from sets SEMA and EXPR.

For the US election dataset (Figure 6.6), the representative models of clusters 1 and 2 have a single feature, domain (a categorical feature) and the relative position of news domain on the Alexa ranking, respectively. These plots are very similar, as there is a close mapping from domains to relative position of them on the Alexa ranking. We found that some domains have a large negative impact (i.e., less likely fake) on the output value such as politico.com, cnn.com and abcnews.go.com. Also, as expected, very low values (top of the Alexa ranking) tend to have a large negative impact on the output. For cluster 3's model, high share counts have large impact – positive or negative – over predictions tending to increase output values. This is consistent with recent research that shows that fake news are more likely to be shared [139, 277]. Cluster 4's model also includes number of shares as a feature and can be interpreted in the same way. Additionally, it includes categories of political bias as a feature. As expected, category has a negative impact on output for mainstream news (purple dots), but increases the prediction value if the source exhibits political bias. Last, the representative models of clusters 5 and 6 include number of shares and reactions as a feature. Once again, we note that higher values of these engagement features increase the chances that a piece of news is classified as fake. Moreover, these clusters include as a feature information regarding source (i.e., page as a categorical feature in cluster 6, and user engagement metrics of Facebook pages that published news stories – page talking about count – as a feature in cluster 5). In cluster 5's model, low values for 'page talking about count' feature are almost always associated with both negative and positive impact, since very well known pages are less likely to share fake news. Last, the plot in cluster 6 is very similar with cluster 1 plot showing that some pages have a large negative impact (i.e., less likely fake) on the output.

The representative models of all clusters for the Brazilian Election dataset (Figure 6.7) have a feature regarding external propagation measures, i.e., information about the dissemination of the news stories on the Web Cluster 1's model includes a single feature from this category: the number of websites/domains that published this news story over the Web. We observe that while high domains count have positive impact over predictions, low values have a negative impact over them tending to increase/decrease output values. Differently, in the cluster 2's model, which also includes the same feature included in cluster 1's model, high domains count that published the news stories over the Web have large impact – positive or negative – over predictions tending to increase output values. Representative models of clusters 3 e 4 also include the number of domains that published the news stories over the Web as a feature and can be interpreted in the same way (clusters 1 and 2, respectively). Additionally, these representative clusters' models include as a feature semantic information regarding text that are almost always associated with positive impact on outputs, highlighting that the fake news disseminated during the 2018 Brazilian election has a strong connection with specific facts.

Last, in Figures 6.8 and 6.9 we present the SHAP results for the top performing models w.r.t AUC for the US and Brazilian election datasets, respectively. Differently from the models closest to the centroids that have one or two features, the clusters in Figures 6.8 and 6.9 have much more features. Figure 6.8 for the US Election dataset shows that all clusters rank engagement features (i.e., number of shares and reactions) nearly as the most impactful features. Cluster 1 uses as a features information regarding



(a) Cluster 1

(b) Cluster 2

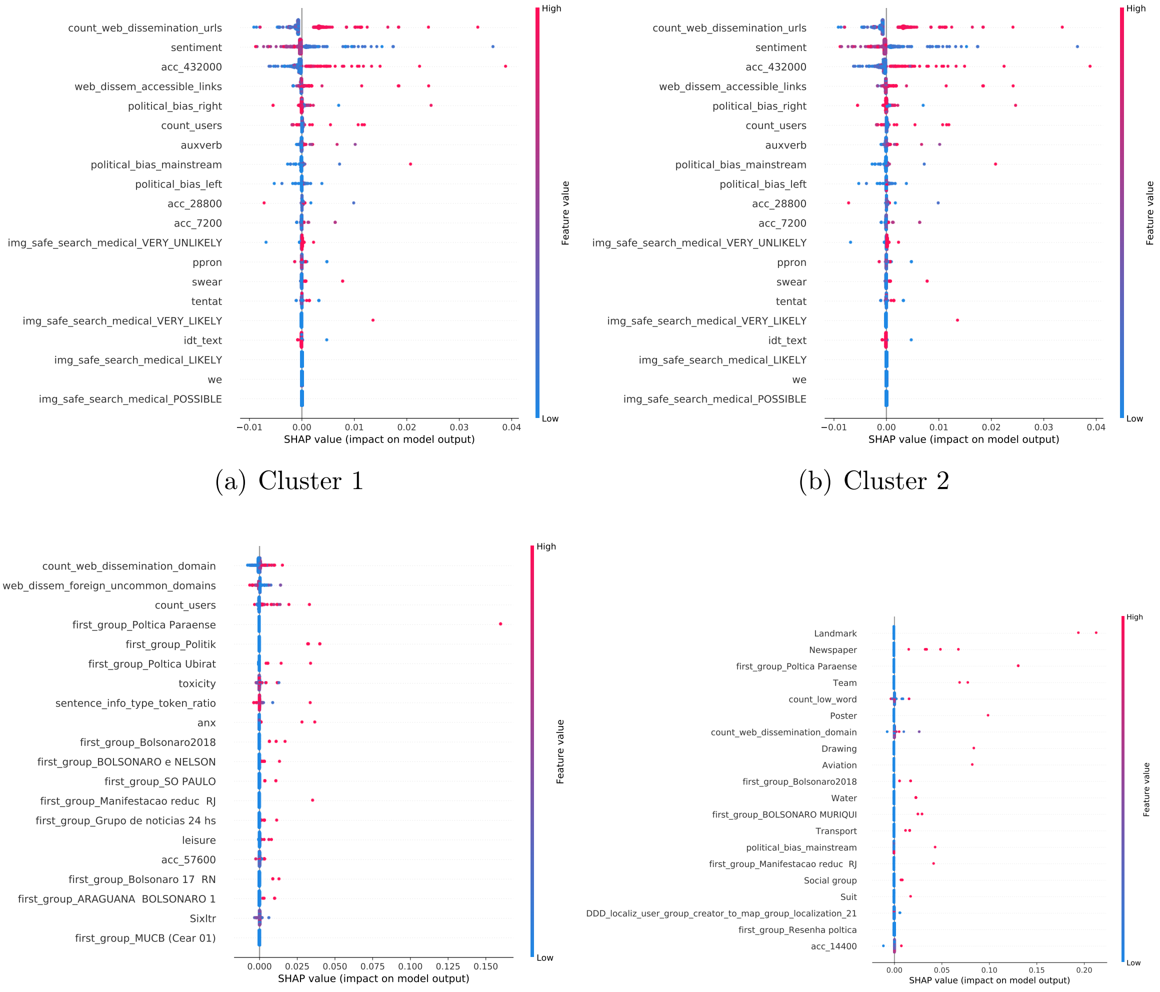
(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Figure 6.8: SHAP summaries for the highest AUC model in each cluster for the US election dataset.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

Figure 6.9: SHAP summaries for the highest AUC model in each cluster for the Brazilian election dataset.

news publisher and indicators of text quality, which are features proposed in the present work. Clusters 2 and 3 use localization (i.e., longitude and IP address/ASN) and news publisher features, including aspects of credibility (or popularity) and trustworthiness of them. For cluster 4, Alexa's ranking position appears as the most important feature, similar to the centroid clusters. It also uses as a features domain localization, image properties, and semantic aspects of a text. Cluster 5 and Cluster 6 are a mix of political bias, information regarding news publisher, external propagation measures, and image properties features, which are top-of-the-rank features on the others clusters. Psycholinguistic cues are shown to be relevant in all clusters, once they appear in every model. For the Brazilian Election dataset (Figure 6.9), all cluster are also a mix of political bias, information regarding news publisher, external propagation measures, semantic aspects of a news story, and image properties features. Similar findings were obtained when analyzing the lowest variability models in each cluster for both datasets, where most of them contain Psycholinguistic cues, News Publishers, Semantic and Engagement features.

6.5 Pareto-Efficient Analysis of Features for Fake News Detection in Different Scenarios

After understanding the general informativeness of the features for fake news detection, we turn our focus to contrast them in different scenarios. Specifically, in this section we investigate if *there is a set of features that yield models with high performance able to identify fake news disseminated on digital platforms considering different scenarios*, i.e., the 2016 US and 2018 Brazilian presidential elections. To accomplish this goal, we propose an experiment based on Pareto-Efficiency which is a central concept in Economics widely explored in several areas of knowledge, including computer science [150, 295].

The Pareto-Efficiency concept states that "when some action could be done to make at least one person better off without hurting anyone else, then it should be done". This action is called Pareto improvement, and a system is said to be Pareto-Efficient if no such improvement is possible [223].

Thereby, the same concept may be exploited to investigate if there is a set of features that can be useful to detect fake news in different scenarios. In this case, the most efficient set of features is the one that cannot improve an objective further (e.g., models with high performance in terms of AUC in a given dataset), without hurting the other objective (i.e., models with high performance learned in another dataset). In other words, it allows to discover the features for fake news detection that generate models with high performance considering data from elections in different countries: the United States and Brazil.

Thus, to perform this analysis, we first filter among those features presented in Chapter 4, only the ones computed in both analyzed datasets. That is, we discard all features extracted from a specific dataset (e.g., information regarding source credibility that are extracted specifically from the US election dataset – and, in the same way, the WhatsApp groups where the news stories were disseminated, which are features extracted specifically from the Brazilian election dataset). In total, there are 163 features for fake news detection that are common to both datasets. These features were used as input for a new round of generation of unbiased models according to our strategy proposed in the Section 6.1 (**Step 1**). In total, we generated 247,941 feature sets (maximum of 20 features in each set) which are used to build models using data from the US and Brazilian elections. Here, our interest is to understand features for fake news detection that can be useful to build models with high performance in terms of AUC considering data from different scenarios.

As shown in Figure 6.10, each possible feature set is associated with a point in a bi-dimensional scattergram (which we call the models space). In this case, a point is represented as [x,y], where each coordinate x, y corresponds to the performance of models in terms of AUC in each of the scenarios analyzed. In total, there are 247,941 points in the graph that correspond to the total of features set generated from unbiased model generation strategy. The blue dots connected by the red line indicate the optimal choice in terms of the feature set, forming a Pareto boundary under the remaining choice space (below and left, gray dots). These points are not dominated by any other point in the scattergram [193, 298], that is, they correspond to cases for which no Pareto improvement is possible, being, therefore, feature sets more likely to be simultaneously accurate to built models with high performance considering data from the US and Brazilian elections. Hence, to obtain a deep understanding of these features that can be useful for detecting fake news in different scenarios, we focus on the 14 models (i.e., sets of features) that compose the Pareto boundary.

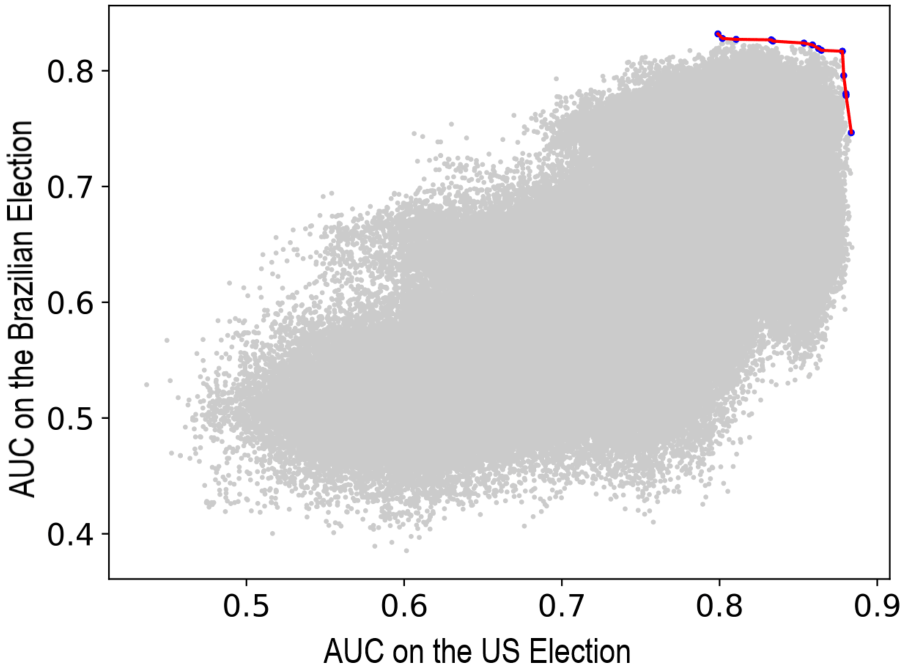


Figure 6.10: Pareto efficiency.

First, observe that the AUC is in the range $[0.80, 0.88]$ for the US election dataset and $[0.75, 0.83]$ for the Brazilian election dataset. Also, we note that among models that compose the Pareto boundary, 28% of them (that is, 4 models) achieved performance less that 0.8 in terms of AUC. Overall, the models have a good trade-off between performance of models from both datasets.

Then, to better understand the relationship between features and model performance in each of the datasets, we compute the prevalence of features. The results are shown in Table 6.1. We observe that features extracted from environment are present in all models that compose the Pareto boundary (e.g., the number of Websites/domains that published the image associated with news story over the Web (100%) and the volume of them that are uncommon (100%)). Moreover, some features extracted from content are very frequent (e.g., identifier of the existence of endorsement text (55%) and count of uppercase words (55%)) in these models. Last, other set of features from source (e.g., information regarding bias of news publisher (44%)) and content, including images properties (e.g., count of image objects (44%)), temporal patterns (44%) and information regarding semantic structure of content (44%), are quite prevalent in this group of models. We conclude that there are features that yield models with high performance in different scenarios.

Table 6.1: Top-10 features in models that compose the Pareto boundary.

Feature	(%)
count_web_dissemination_urls (env:EXPR)	100
web_dissem_foreign_uncommon_domains (env:EXPR)	100
idt_text (content:SYNT)	55
count_up_word (content:LEXI)	55
acc_432000 (env:TEMP)	44
img_objects (content:SEMA)	44
word_usage_pronoun (content:LEXI)	44
img_count_objects (content:IMAG)	44
focuspast (content:PSYC)	44
political_bias (source:BIAS)	44

6.6 Summary

Based on our survey on features for fake news detection as well as our attempt to implement all potential features to detect fake news, and our experimental results described in the previous chapters of this thesis, in this chapter, we explored our RG2 whose objective is to investigate the informativeness of features for fake news detection providing explanations about the decisions made by these algorithms designed for this task. Particularly, we presented our unbiased framework for quantifying the informativeness of features for fake news detection. Our results unveil the real impact of a slew of features for fake news detection considering data from two specific scenarios: the US and Brazilian elections. First, our proposed features for fake news detection related to the news source appear within the best models up to five times more often than other features. Second, our framework demonstrates how hard is to detect fake news, as only a small fraction of the models achieved a good detection performance in terms of AUC.

In addition, our findings suggest that fake news with specific patterns tend to be identified by models with specific combinations of features. As result, different models separate fake stories from real ones based on very different reasoning. This shows the complexity of the problem and allows us to understand how hard it is for a single solution to tackle all forms of fake news stories. Further, as part of our proposed framework, we presented explanations of what features or set of them tend to play major roles in classifying a news story as fake for both datasets. Explaining predictions made by models designed for fake news detection is crucial and can be useful to help fact-checkers understand clearly models decisions, supporting them on the fact-checking process. For example, a certain story was considered as false because it was posted by new newspaper hosted under the same IP address (ASN) as a known blacklisted fake news source. Moreover, we present a set of features that are useful for detecting fake news in different scenarios.

Last, the next chapter explores strategies with practical potential for detecting fake news.

Chapter 7

Fakeness Score Model

In this thesis, we discuss that automatic solutions for fake news detection could be used as an assistive tool for fact-checkers to identify content that is more likely to be fake or content that is worth checking, still leaving the final call to an expert at the endpoint of the process. Thus, we leverage our findings towards building automatic detection tools to incorporate them into a real system to help in the fact-checking task. Particularly, we incorporate our approach in the WhatsApp Monitor, a system to support fact-checkers available at the following link: `http://www.whatsapp-monitor.dcc.ufmg.br/`.

The WhatsApp Monitor is a web-based system proposed by our research group [167, 217] that provides a way to explore the most popular content shared on WhatsApp public groups and displaying them daily in an organized way. Such an organization has been used by many journalists and agencies, including Comprova, a collaborative journalistic project from First Draft focused on verifying questionable stories published on digital platforms and WhatsApp during the 12 weeks leading up to the Brazilian 2018 presidential election.

This system integrates a framework that (i) collects data from hundreds of public groups on WhatsApp, monitoring multiple types of media such as images, videos, audio, and textual messages from different countries, including Brazil, India and Indonesia; (ii) matches identical pieces of information together and ranks them every day; (iii) and, finally, displays the ranked content on a web application where users can navigate through days and content. Although this system has already been used extensively, it only displays a list of the most frequently shared content in the monitored groups over a time interval. This strategy did not necessarily indicate which content should be fact-checked first, as the popularity of a news story in WhatsApp may not be representative of its popularity elsewhere.

Therefore, based on the framework presented in this thesis, we propose and implement a new ranking mechanism that accounts for the potential occurrence of fake news within the data, significantly reducing the number of content pieces journalists and fact-checkers have to go through before finding a fake story. Particularly, we explore the XGB machine learning method to estimate a *fakeness score* on news stories aiming at improving ranking results, which can support decisions regarding the selection of facts (or news) to be checked. Last, we deploy our approach in a real system, the WhatsApp Monitor.

Our experimental evaluation shows that this tool can reduce by up to 40% the amount of effort required to identify 80% of the fake news in the data when compared to current mechanisms practiced by the fact-checking agencies for the selection of material to be checked such as popularity ranking.

7.1 Problem Statement

From a technical standpoint, we address the fake news detection as a ranking problem: given a set S of news stories disseminated through images on WhatsApp, its goal is to generate a sorted list R in which news stories that are more likely to be fake are displayed at the top positions of the list. The architecture of the proposed ranking tool is organized into the modules shown in Figure 7.1.

We use the new dataset built in Chapter 3 (WhatsApp Data), containing news stories disseminated during the 2018 Brazilian presidential election verified by expert fact-checkers (Labeled Data) to train a model that outputs a fakeness score. In order to do so, we explore our features for fake news detection presented in Chapter 4 (Feature Extraction) to train a machine learning model (Fakeness Scoring). This model is used to

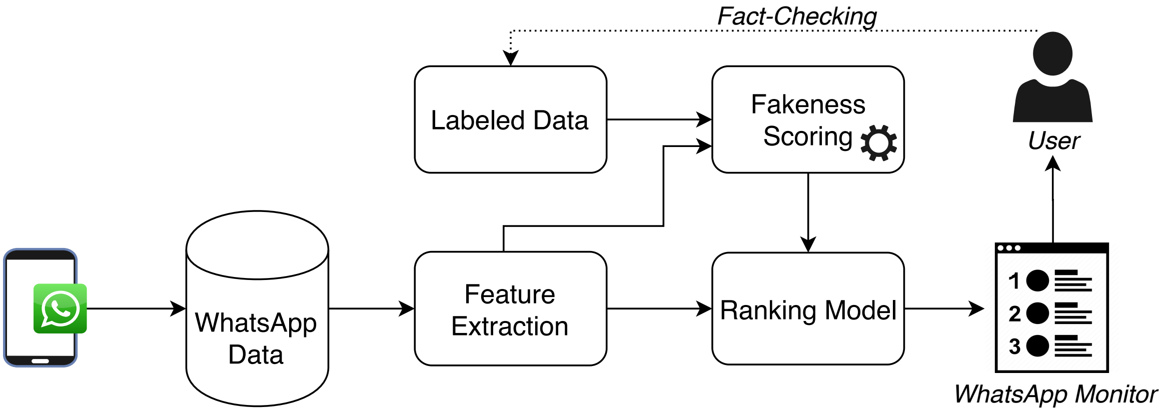


Figure 7.1: Architecture of the proposed tool for ranking WhatsApp content w.r.t. fakeness scores.

compute scores based on extracted features for new content collected from WhatsApp, and then to rank them (Ranking Model). Last, we make this ranking available in the WhatsApp Monitor for users, which can include fact-checkers that may eventually help us improve the model by providing more labeled data.

7.2 Experimental Evaluation

In this section, we describe our ranking-based strategy and the details of our experimental setup, including the metrics used to assess the performance of the methods. At the end, we present and discuss the main results.

7.2.1 Ranking-Based Strategy

Based on results presented in Chapter 5 of this thesis, we explore the XGB classifier. Specifically, the XGB model "learn" from training data how to assign a fakeness score to a news story disseminated through images on WhatsApp. Each instance i of the training dataset is composed by a feature vector X_i, containing the values of the features described in Chapter 4, and a label y_i indicating whether i is fake (y_i=1) or unchecked (y_i=0). Given an unseen news story j, the output is an estimate of the probability of j being fake, which in turn is used to produce a ranking of news stories.

7.2.2 Evaluation Metrics

Since our goal is to build good rankings, we assess the effectiveness of our ranking-based strategies using metrics commonly adopted in Information Retrieval: Precision, Recall and Normalized Discounted Cumulative Gain (NDCG) [21] in the top k positions of the ranking (P@k, Recall@k and NDCG@k, respectively) for different values of k. In the fact-checking domain, k represents the number of news that the fact checker specialist can afford to inspect.

Let S be the set of all news to be checked, $F \subseteq S$ be the set of fake news among them, R be the produced ranking, and R^k be the top-k positions of R. Precision@k is the fraction of fake news detected in the first k positions of the provided rank, that is:

$$Precision@k(R, F) = \frac{|R^k \cap F|}{k}. \tag{7.1}$$

Recall@k is the fraction of all existing fake news in S that were indeed retrieved among the top-k positions of the ranking:

$$Recall@k(R,F) = \frac{|R^k \cap F|}{|F|} \tag{7.2}$$

The NDCG@k metric, in turn, emphasizes results in the top positions of the ranking. Let $DCG@k$ be the discounted cumulative gain in the first k positions of the ranking, defined below in the equation 7.3, where $f(i)$ is equal to 1 if the i-th news story in R is fake (i.e., it is in F), and 0 otherwise. The un-normalized and normalized discounted cumulative gain in the first k recommendations are respectively:

$$DCG@k(R,F) = \sum_{i=1}^{k} \frac{f(i)}{\log_2(i+1)}, \tag{7.3}$$

$$NDCG@k(R,F) = \frac{DCG@k(R,F)}{IdealDCG@k}, \tag{7.4}$$

where $IdealDCG$ is the value obtained for $DCG@k$ when there are only fake news at the top-k (or fewer) positions.

7.2.3 Experimental Setup

We evaluate the effectiveness of the fake news ranking methods using 5-fold cross-validation experiments. That is, the set of news stories is partitioned into five equal-sized (stratiefied) folds. Three folds are used as training set to "learn" the models. One fold is used as validation set, for parameter tuning, and the last fold is used for testing. The folds are rotated such that all possible 5 configurations of training, validation and test sets are tested.

In order to obtain multiple lists of news to rank, for each portion in each fold, we generated 50 bootstrap samples with replacement (also stratified), containing 200 news each. Thus, reported results are averages over 50 samples × 5 folds = 250 executions.

7.3 Results

Table 7.1 shows average Precision@k, Recall@k and NDCG@k results (for k=5, 10, 50 and 100) yielded by each strategy, namely, a baseline in which news are ranked by their number of shares ($\#Shares_{Rank}$), and the ranking-based technique XGB. Also,

Table 7.1: Average experimental results and 95% confidence intervals. Best results (and statistical ties) in bold.

	Precision@5	Precision@10	Precision@50	Precision@100
$\#Shares_{Rank}$	0.092 ± 0.020	0.105 ± 0.014	0.064 ± 0.004	0.045 ± 0.002
XGB	**0.237** ± 0.025	**0.173** ± 0.016	**0.085** ± 0.003	**0.053** ± 0.001
	Recall@5	Recall@10	Recall@50	Recall@100
$\#Shares_{Rank}$	0.077 ± 0.017	0.176 ± 0.023	0.530 ± 0.033	0.744 ± 0.031
XGB	**0.198** ± 0.021	**0.288** ± 0.027	**0.706** ± 0.027	**0.891** ± 0.020
	NDCG@5	NDCG@10	NDCG@50	NDCG@100
$\#Shares_{Rank}$	0.110 ± 0.025	0.113 ± 0.017	0.076 ± 0.006	0.057 ± 0.003
XGB	**0.250** ± 0.029	**0.201** ± 0.020	**0.113** ± 0.006	**0.078** ± 0.004

95% confidence intervals are presented, and best results (with corresponding statistical ties according to a 2-sided t-test) are emphasized in bold.

Our first observation is that, as expected, ranking news stories by the number of shares is not ideal for finding fake news, performing poorly in comparison to the our strategy, w.r.t. all evaluation metrics. Our ranking-based strategy (XGB) outperforms this popularity-based baseline with gains of up to 64%, 64% and 77% in Precision@10, Recall@10 and NDCG@10, respectively. Thus, although the number of shares is a good evidence of the impact that fake news may cause, it is not the best strategy to find fake news earlier in the ranking.

Thus, our results show that our features can be effectively used for learning a good fakeness function being XGB a best performing strategy in comparison to the popularity-based baseline considering all evaluation metrics. In the next section, we analyze our results under the lens of fact-checkers, which leverage our methods to detect fake news earlier.

7.4 Potential Applications

In this section, we discuss potential applications of the *fakeness score* model in terms of cost of fact-checking (Section 7.4.1) and deploy the ranking functionality based on this model in a real application (Section 7.4.2).

7.4.1 Fact-Checking (Cost Analysis)

One of the most important applications of the tool we propose here is the fact-checking process. Ranking images in the WhatsApp Monitor according to a fakeness score can be used to help an expert fact-checker assign priorities in a more informed way.

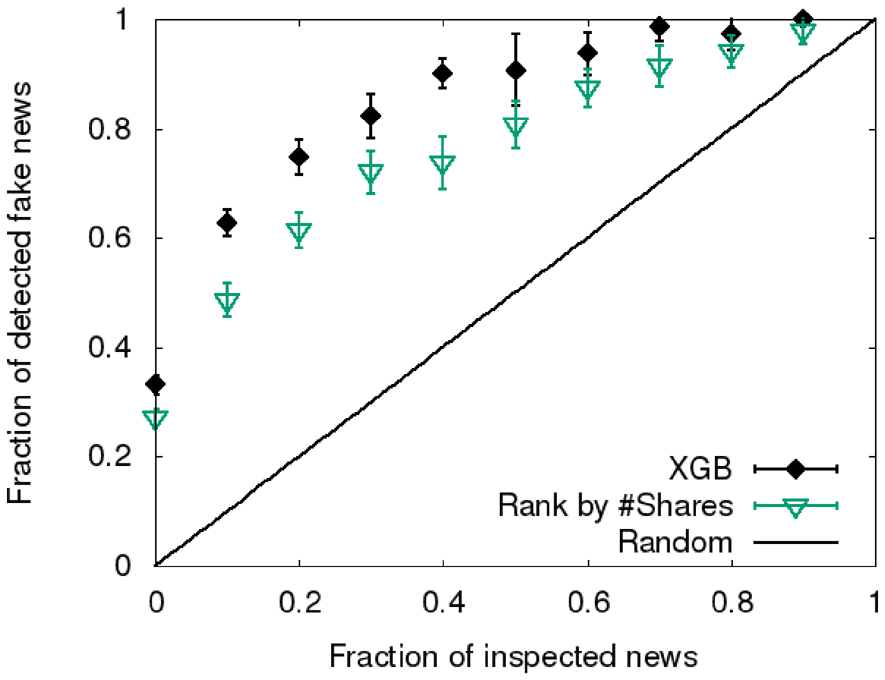


Figure 7.2: Cost analysis.

We now revisit the previous results but from the fact-checker perspective. More precisely, we evaluate, for each strategy (i.e., our ranking-based strategy and popularity-based baseline), (i) the effort required to recover a given fraction of the fake news and, conversely, (ii) the fraction of fake news recovered upon fact-checking a given fraction of the images. We assume that the fact-checker strictly follows the ranking returned by each strategy and that the cost of fact-checking any image is a fixed constant.

Figure 7.2 shows, for both strategies, the fraction of news inspected (x-axis) and the average fraction of detected fake news with confidence intervals (y-axis). We observe that to recover 80% of the fake news, a ranking created with XGB would require a fact-checker to check approximately 30% of the news, while the popularity-based strategy would require around 50%. Interestingly, by checking the top 40% entries ranked by XGB, the fact-checker would recover roughly 90% of all fake news in the dataset. These results reinforce that the use of machine learning based methods for estimating a fakeness score can significantly reduce the efforts required to identify fake news, potentially allowing them to be found at an early stage.

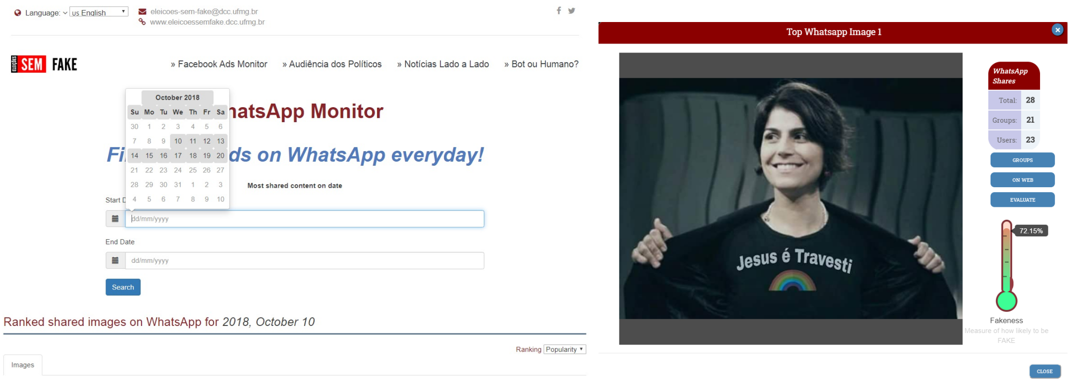
7.4.2 WhatsApp Monitor

Finally, to test our approach, we deployed the ranking functionality developed by our methodology in a real application for exhibiting WhatsApp data, thus extending the system proposed in [167]. Their work provides an online system where users can oversee the daily trends shared on the WhatsApp public groups for a particular country (e.g., Brazil, Indonesia and India) or domain (e.g., politics and news). The system ingests a large amount of images from the groups being monitored and, then, shows popular pieces of media content related to the political and news topics.

WhatsApp was pointed as one of the main sources of fake news in Brazil and India during the last elections [17, 168, 215, 216], however, due to its closed nature, the social structure of this messaging app still remains barely explored. Even though results show that looking into the most shared content is a much better baseline compared to a simple random approach, it is not easy to determine what are the popular media being shared on WhatsApp as it is on Twitter or Facebook. Some works seek to unveil these WhatsApp networks and recover widely shared content within the platform. For instance, Moreno et al. [173] uses WhatsApp for monitoring and responding during critical events in Ghana and, markedly, the WhatsApp Monitor [167].With our proposed extension, which allows users to rank images and prioritize those most likely to be fake in this system, we hope our tool can further assist fact-checking agencies to fight misinformation.

Figure 7.3 presents some screenshots of the WhatsApp Monitor interface[1]. After accessing the tool and logging in, the users of the system are taken to a dashboard where they can navigate between dates and observe the most shared multimedia content in the monitored groups for a given date, as shown in Figure 7.3(a). With the new functionality, users can choose between different methods for ranking the content: popularity, which sorts images by the total number of shares; or by *fakeness*, which is based on the probability that the image is fake, as estimated by our XGB model. The system also has two other ranking methods based on the number of different groups or different users who shared that image. This allows journalists to determine the content shared in WhatsApp that may be worth checking each day. Also, using this new tool, any ranking model can be used as an “off-the-shelf” strategy. That means, different ranking models can be trained and aggregated to the system (e.g., a model trained with pornography or specific fake news related to health), and it is up to the users to choose how to order the content based on their preference.

[1] A working demo of our developed tool can be accessed in `http://blackbird.dcc.ufmg.br/test_monitor/`



(a) Navigation search

(b) Image details

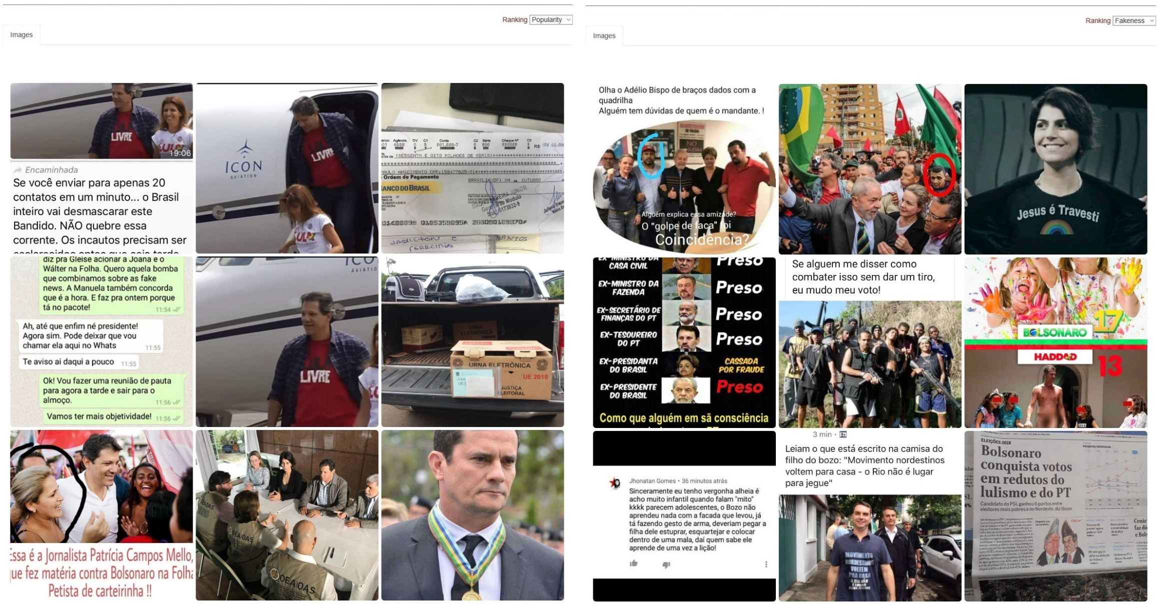
Figure 7.3: Screenshots of WhatsApp Monitor interface.

By clicking on “Details” of an image (Figure 7.3(b)), it shows the number of shares, users, and groups in which the image appears, to help them identify some context associated with the content, and a visual representation of the *fakeness score* as a thermometer, accompanied by the probability value assigned by the trained model. To ensure the privacy of users, we do not share or disclose any Personally Identifiable Information (PII) such as cellphone numbers. We use only ids in order to measure aggregate spreading statistics. Also, to avoid any misuse of the system, we limit the users’ access through a login account. Since we only use publicly available WhatsApp groups we joined, our data collection does not violate WhatsApp terms of service.

Figure 7.4 shows how images are displayed on the system, contrasting the images shown when ranking by popularity and fakeness. We observe that the methods of sorting images generate very distinct views of data. We believe that both rankings can help a journalist. The fakeness ranking (Figure 7.4(b)) presents fake news in all top items in this example, while popular ranking (Figure 7.4(a)) show most shared content for that period, but 5 of the images exhibited there were not verified to be fake. Note that, although the images displayed in the figure and the data used to train our model come from the same source, we picked a different date range than the on used for training the model to ensure a fair comparison.

7.5 Summary

In this chapter, we consider the problem of fake news dissemination through WhatsApp images by investigating how to design and integrate a tool to WhatsApp Monitor system that allows users to rank those images according to an estimated *fakeness score*. Our experimental evaluation shows that our approach provides substantial gains (up


Ranking Popularity
Images
LIVRE
19:06
ICON AVIATION
Ordem de Pagamento
BANCO DO BRASIL
(SESSENTA E OITO MILHOES DE REAIS)
URNA ELETRÔNICA
UE 2015
JUSTIÇA ELEITORAL
Essa é a Jornalista Patrícia Campos Mello,
que fez matéria contra Bolsonaro na Folha
Petista de carteirinha !!
OEA/OAS
Ranking Fakeness
Images
Alguém explica essa amizade?
O "golpe de faca" foi
Coincidência?
Jesus é Travesti
EX-MINISTRO DA CASA CIVIL Preso
EX-MINISTRO DA FAZENDA Preso
EX-SECRETÁRIO DE FINANÇAS DO PT Preso
EX-TESOUREIRO DO PT Preso
EX-PRESIDANTA DO BRASIL CASSADA POR FRAUDE
EX-PRESIDENTE DO BRASIL Preso
Como que alguém em sã consciência
BOLSONARO 17
HADDAD 13
MOVIMENTO NORDESTINOS VOLTEM PRA CASA O RIO NÃO É LUGAR PARA JEGUE
ELEIÇÕES 2018
Bolsonaro conquista votos em redutos do lulismo e do PT
Candidato do PSL ganhou 6 pontos entre eleitores mais pobres e no Nordeste, diz Ibope
PESQUISA DATAFOLHA

Encaminhada
Se você enviar para apenas 20 contatos em um minuto... o Brasil inteiro vai desmascarar este Bandido. NÃO quebre essa corrente. Os incautos precisam ser

diz pra Gleise acionar a Joana e o Wálter na Folha. Quero aquela bomba que combinamos sobre as fake news. A Manuela também concorda que é a hora. E faz pra ontem porque tá no pacote! 11:54

Ah, até que enfim né presidente! Agora sim. Pode deixar que vou chamar ela aqui no Whats 11:55

Te aviso ai daqui a pouco 11:55

Ok! Vou fazer uma reunião de pauta para agora a tarde e sair para o almoço. 11:56

Vamos ter mais objetividade! 11:56

Olha o Adélio Bispo de braços dados com a quadrilha
Alguém tem dúvidas de quem é o mandante. !

Se alguem me disser como combater isso sem dar um tiro, eu mudo meu voto!

Jhonatan Gomes • 36 minutos atrás
Sinceramente eu tenho vergonha alheia é acho muito infantil quando falam "mito" kkkk parecem adolescentes, o Bozo não aprendeu nada com a facada que levou, já tá fazendo gesto de arma, deveriam pegar a filha dele estuprar, esquartejar e colocar dentro de uma mala, daí quem sabe ele aprende de uma vez a lição!

3 min ·
Leiam o que está escrito na camisa do filho do bozo: "Movimento nordestinos voltem para casa - o Rio não é lugar para jegue"

(a) Popularity ranking

(b) Fakeness score ranking

Figure 7.4: Screenshots of the WhatsApp Monitor comparing the different ranking strategies for the same period.

to 27%) in terms of precision and recall when compared to the baseline, considerably improving current mechanisms adopted by the checking agencies for selecting the news stories to be checked such as popularity ranking. Then, we discussed potential applications of this tool to fact-checking. In our experiments, the proposed tool reduced by up to 40% the amount of effort required to identify 80% of the fake news, hence significantly contributing to the fact-checking process. Moreover, we validate our approach by integrating the *fakeness score* model to a real system extensively used by Brazilian fact-checking agencies.

The next section presents a summary of the main results of this thesis and provides directions for future work.

Chapter 8

Conclusions

In this chapter, we summarize the main contributions of this thesis (Section 8.1). Moreover, we also present a discussion on future research directions (Section 8.2), as well as the list of publications derived from this thesis (Section 8.3).

8.1 Summary of Results

In this thesis, we studied the practical potential of automatic solutions to identify fake news disseminated on digital platforms. Particularly, (i) we surveyed datasets and features for fake news detection, and then, we explored the prediction performance of current approaches and features for automatic detection of fake news in different scenarios. (ii) We also developed a framework for quantifying the real informativeness of features for fake news detection, and last (iii) we proposed a *fakeness score* model as a way to help fact-checking agencies identify fake news stories shared through images on WhatsApp. Although there are some isolated initiatives that study fake news spread in Brazil [172], to the best of our knowledge, this is the first work that explores strategies with practical potential for detecting fake news spread on WhatsApp. Such studies are categorized in research goals, which are summarized in Sections 8.1.1 to 8.1.3.

8.1.1 RG1 - Assessing the Prediction Performance of Solutions to Detect Fake News

We first surveyed a large number of recent and related works as an attempt to implement all potential features to detect fake news. In this field, we explored data from different scenarios: the 2016 US Election and Health. We also built a new dataset containing news stories disseminated through images on WhatApp during the 2018

Brazilian presidential election (Chapter 3). We believe that this dataset can be useful for research in a variety of contexts such as better comprehension of the fake news landscape in Brazil and the development of new automatic detection tools.

Moreover, we also proposed 22 new features for fake news detection (Chapter 4), such as those related to the news source (e.g., IP address, domain, information regarding source credibility), an indicator of toxicity of text, image safe search indicators, external propagation measure (outside digital platforms) and readability features to assess the writing style of news stories, which some of them (e.g., features from news source) appear within the best models more often than other features from previous efforts.

We then evaluated and compared different supervised machine learning approaches, assessing their prediction performance in the task of automatically identify fake news disseminated in different scenarios (Chapter 5). Our results revealed that automatic fake news detection could be used by fact-checkers as an auxiliary tool for identifying content that is more likely to be fake. Particularly, we showed that the prediction performance of proposed features combined with existing classifiers has a useful degree of discriminative power for detecting fake news. Our best classification results can correctly detect nearly all fake news in our data while misclassifying about 40% of true news for the US election dataset and 55% of the unchecked content for the Brazilian election dataset, which is already sufficient to help fact-checkers. In contrast, the results also showed that our proposed features and models for fake news detection were not useful in correctly distinguishing health fake news from others, suggesting that fake news detection is a context-dependent task.

In summary, our study on RG1 offers an initial step towards automatic fake news detection disseminated on digital platforms. We hope it motivates follow-up efforts covering other datasets, scenarios/contexts, time periods, languages. Also, by sharing our new dataset built in this thesis, we hope other researchers can provide other perspectives of understanding of the fake news phenomena, triggering new countermeasures in the upcoming elections.

8.1.2 RG2 - Quantifying the Informativeness of Features for Fake News Detection

We proposed a framework for quantifying the informativeness of features for fake news detection which revealed how hard is to detect fake news, as only a small fraction of the models achieved a good detection performance in terms of AUC (Chapter 6). We hope our effort can become a baseline for other solutions to the same problem.

Our findings also suggested that fake news with different patterns tend to be identified by models with specific combinations of features. As result, different models separate fake stories from real/unchecked ones based on very different reasoning. This shows the complexity of the problem and allows us to understand how hard it is for a single solution to tackle all forms of fake news stories in different scenarios. As future work, we plan to categorize the fake news stories as a strategy to construct effective and robust ensembles of classifiers. For instance, in this work, we showed the different models of clusters that are made of random combinations of features.

A clever way of increasing the chances of performance enhancement in a classifier is by making an ensemble out of the different models from the clusters that we found, once we combine their features, the probability of detecting a piece of fake news might be higher. This indicates that ensemble techniques that combine models from different clusters are a promising avenue for investigation.

8.1.3 RG3 - Exploring the Practical Potential of Fake News Detection

We considered the problem of fake news dissemination through WhatsApp images by investigating how to design and integrate a tool to our WhatsApp Monitor system [167] that allows users to rank those images according to an estimated *fakeness score* (Chapter 7). To achieve this goal, we explored the new dataset built in this thesis (Chapter 3) containing news stories disseminated during the 2018 Brazilian Election to evaluate the effectiveness of strategies. In this process, we also explored again our features for fake news detection that extract content, source and environment information to fit those models (Chapter 4). Our experimental evaluation showed that our approach provides substantial gains (up to 27%) in terms of precision and recall when compared to the baseline, considerably improving current mechanisms adopted by the checking agencies for selecting the news stories to be checked such as popularity ranking.

We discussed potential applications of this tool to fact-checking. In our experiments, the proposed tool reduced by up to 40% the amount of effort required to identify 80% of the fake news, hence significantly contributing to the fact-checking process. Moreover, we validate our approach by integrating the fakeness score model to a real system extensively used by Brazilian fact-checking agencies. As future work, we intend to conduct A/B tests [73] to evaluate the effectiveness of the rankings generated by our approach in practice.

Moreover, given the sheer volume and heterogeneity of data from WhatsApp w.r.t. content, source, etc, we plan to investigate whether it is possible to generate good rankings from reduced subsets of features in order to speed up the feature extraction process or even to improve the results by increasing the signal to noise ratio. Also, providing explanations that supported the algorithm's output is crucial in this context. Thus, we intend to explore strategies to enhance the interpretability of our models incorporating them into WhatsApp Monitor, as a way to help fact-checkers understanding the influence of each of the features on the decisions of the algorithms, supporting their checks.

Last, our proposed approach requires a continual pipeline where more stories get labeled each day and are, in turn, fed back to the models. Rather than verifying only the most suspicious stories, an active learning solution can be put in place, so that the model can also indicate which stories should be investigated in order to improve its prediction performance.

8.2 Future Research Directions

Despite the importance of the results achieved in this thesis, including contributions provided by concurrent efforts, countering fake news is a typical adversarial issue that requires continuous studies. Every election, misinformation campaigns explore new ways to manipulate opinion and new defense mechanisms are created aiming at least to mitigate the misinformation campaigns. As an example, the efforts that attempt to understand the abuse of WhatsApp in the Brazilian elections have motivated countermeasures deployed in the Spanish elections[1,2] as well as worldwide changes in WhatsApp to slow down the dissemination of viral content[3].

Thus, this thesis opens an avenue of directions for future work in a multidisciplinary field where journalists and computer scientists can explore together new features (e.g., alternative media identifiers [104], additional lexicons [6] and network-based metrics [230, 305]) and new combinations of them, more sophisticated learning models (e.g., using active learning approaches), and data from other scenarios (e.g., COVID-19 [59], wars [232]) to design new automatic mechanisms for detecting fake news on digital platforms that are increasingly effective. Moreover, larger volumes of labeled data will enable to explore, in the future, other techniques such as deep learning and push the boundaries of prediction performance.

[1] https://elpais.com/elpais/2019/03/18/inenglish/155290078_672737.html

[2] https://www.independent.co.uk/news/world/europe/spain-elections-whatsapp-podemos-channel-close-left-ing-de-olmo-a8886481.html

[3] https://www.theguardian.com/technology/2019/jan/21/whatsapp-limits-message-forwarding-fight-fake-news

Finally, another interesting topic related to machine learning that is worth further studying in our problem is the transfer learning from models learned in a given dataset and applied to another. In machine learning tasks, transfer learning occurs when an algorithm uses knowledge obtained while solving a specific problem (source-task) and applying it to a different but related one (target-task) [166]. This concept is also used as a way to offset the difficulties posed by tasks that involve unsupervised learning, semi-supervised learning, or small datasets [266]. Thus, evaluating whether it is possible to explore some "easily available" knowledge from an external source by analyzing our datasets and models from different scenarios is also a promising future work.

8.3 Bibliographical Contributions

The main results of this thesis generated the following publications:

- REIS, J. C. S.; Melo, P.; Belém, F.; Murai, F.; Almeida, J.; Benevenuto, F. Helping Fact-Checkers Identify Fake News Stories Shared through Images on WhatsApp. In *Information Systems Journal Special Issue on Misinformation on the Web (under review)*, 2020.

- REIS, J. C. S.; Melo, P.; Garimella, K.; Benevenuto, F. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? In *The Harvard Kennedy School (HKS) Misinformation Review*, 2020.

- REIS, J. C. S.; Melo, P.; Garimella, K.; Almeida, J.; Eckles, D.; Benevenuto, F. A Dataset of Fact-Checked Stories Shared in WhatsApp during Brazilian and Indian Elections. In *Proc. of the Int'l AAAI Conference on Web and Social Media (ICWSM)*, 2020.

- REIS, J. C. S.; Correia, A.; Murai, F.; Veloso, A.; Benevenuto, F. Supervised Learning for Fake News Detection. In *IEEE Intelligent Systems.* Volume 34, Number 2, 2019.

- REIS, J. C. S.; Correia, A.; Murai, F.; Veloso, A.; Benevenuto, F. Explainable Machine Learning for Fake News Detection. In *Proc. of the Int'l ACM Web Science Conference (WebSci)*, 2019.

- REIS, J. C. S.; Kwak, H.; An, J.; Messias, J.; Benevenuto, F. Demographics of News Sharing in the US Twittersphere. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, 2017.

During the development of this thesis, we were also involved in other studies that are indirectly related to its topic and embraced opportunities to collaborate with other researchers. They generated the following publications:

- Lima, L.; REIS, J. C. S.; Melo, P.; Murai, F.; Benevenuto, F. Characterizing (Un)moderated Textual Data in Social Systems. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020.

- Guimarães, S.; REIS, J. C. S.; Lima, L.; Ribeiro, F.; Vasconcelos, M.; An, J.; Kwak, H.; Benevenuto, F. Identifying and Characterizing Alternative News Media on Facebook. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020.

- Guimarães, S.; REIS, J. C. S.; Ribeiro, F.; Benevenuto, F. Characterizing Toxicity on Facebook Comments in Brazil. In *Proc. of the Brazilian Symposium on Multimedia and Web (WebMedia)*, 2020.

- Kansaon, D.; Brandão, M.; REIS, J. C. S.; Barbosa, M.; Matos, B.; Benevenuto, F. Mining Portuguese Comparative Sentences in Online Reviews. In *Proc. of the Brazilian Symposium on Multimedia and Web (WebMedia)*, 2020.

- Resende, G.; Melo, P.; REIS, J. C. S.; Vasconcelos, M.; Almeida, J. M.; Benevenuto, F. Analyzing the Spread of Textual Information in WhatsApp Groups. In *Proc. of the Int'l ACM Web Science Conference (WebScience)*, 2019.

- Lima, L.; REIS, J. C. S.; Melo, P.; Murai, F.; Araujo, L.; Vikatos, P.; Benevenuto, F. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018.

- Christhie, W.; REIS, J. C. S.; Benevenuto, F.; Moro, M. M.; Almeida, V. Detecção de Posicionamento em Tweets sobre Política no Contexto Brasileiro. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2018.

- REIS, J. C. S.; Miranda, M.; Bastos, L.; Prates, R.; Benevenuto, F. Uma Análise do Impacto do Anonimato em Comentários de Notícias Online. *In Proc. of the Brazilian Symposium on Collaborative Systems (SBSC)*, 2016 (*Best Paper Award*).

- Ramos, P.; REIS, J. C. S.; Benevenuto, F. Uma Análise da Polaridade Expressa nas Manchetes de Notícias Brasileiras. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2016.

- Araújo, M.; REIS, J. C. S.; Pereira, A.; Benevenuto, F. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proc. of the Annual ACM Symposium on Applied Computing (SAC)*, 2016.

- REIS, J. C. S.; Gonçalves, P. ; Araújo, M. ; Pereira, A. C. M. ; Benevenuto, F. Uma Abordagem Multilíngue para Análise de Sentimentos. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2015.

- Gonçalves, P.; Hasan, D.; REIS, J. C. S.; Messias, J.; Ribeiro, F.; Melo, P.; Araujo, L.; Benevenuto, F.; Gonçalves, M. Bazinga! Caracterizando e Detectando Sarcasmo e Ironia no Twitter. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2015.

Bibliography

[1] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. In *Proc. of the Int'l Conference on User Modeling, Adaptation, and Personalization (UMAP)*, pages 1--12.

[2] Agarwal, P., Vaithiyanathan, R., Sharma, S., and Shroff, G. (2012). Catching the long-tail: Extracting local news events from twitter. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 379--382.

[3] Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Proc. of the Int'l Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*, pages 127--138.

[4] Alkhodair, S. A., Ding, S. H., Fung, B. C., and Liu, J. (2020). Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management*, 57(2):102018.

[5] Allport, G. W. and Postman, L. (1947). The psychology of rumor. *Henry Holt. American Psychological Association.*

[6] Amorim, E., Cançado, M., and Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies*, pages 229--237.

[7] An, J., Cha, M., Gummadi, K., and Crowcroft, J. (2011). Media landscape in twitter: A world of new conventions and political diversity. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 18--25.

[8] An, J., Cha, M., Gummadi, K. P., Crowcroft, J., and Quercia, D. (2012). Visualizing media bias through twitter. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 2--5.

[9] An, J. and Kwak, H. (2016). Multidimensional analysis of gender and age differences in news consumption. In *Computation and Journalism Symposium (C+J)*.

[10] An, J. and Kwak, H. (2017). What gets media attention and how media attention evolves over time - large-scale empirical evidence from 196 countries. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 464--467.

[11] An, J., Quercia, D., Cha, M., Gummadi, K., and Crowcroft, J. (2014). Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science*, 3(1):12.

[12] An, J. and Weber, I. (2016). # greysanatomy vs.# yankees: Demographics and hashtag use on twitter. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 523--526.

[13] Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490--496.

[14] Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *Proc. of the Network and Distributed System Security Symposium (NDSS)*, pages 1--15.

[15] Armstrong, C. L. and Gao, F. (2011). Gender, twitter and news content: An examination across platforms and coverage areas. *Journalism Studies*, 12(4):490--505.

[16] Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027--1035.

[17] Arun, C. (2019). On whatsapp, rumours, and lynchings. *Economic & Political Weekly*, 54(6):30--35.

[18] Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., and Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1--27.

[19] Babaei, M., Kulshrestha, J., Chakraborty, A., Benevenuto, F., Gummadi, K. P., and Weller, A. (2018). Purple feed: Identifying high consensus news posts on social

media. In *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 10--16.

[20] Badawy, A., Addawood, A., Lerman, K., and Ferrara, E. (2019). Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, 9(1):31.

[21] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[22] Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130--1132.

[23] Bandari, R., Asur, S., and Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 26--33.

[24] Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Proc. of the Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, pages 12--21.

[25] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proc. of the Int'l ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*, pages 49--62.

[26] Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11).

[27] Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proc. of the IEEE Int'l Conference on Big Data (Big Data)*, pages 556--565.

[28] Bhattacharya, D. and Ram, S. (2012). Sharing news articles using 140 characters: A diffusion analysis on twitter. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 966--971.

[29] Blevins, C. and Mullen, L. (2015). Jane, john... leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly*, 9(3):17--35.

[30] Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7.

[31] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

[32] Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal of Communication*, 66(3):343--365.

[33] Budak, C., Goel, S., and Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250--271.

[34] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1301--1309.

[35] Bursztyn, V. S. and Birnbaum, L. (2019). Thousands of small, constant rallies: A large-scale analysis of partisan whatsapp groups. In *Proc. of the Int'l IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 484--488.

[36] Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74--80.

[37] Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[38] Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 675--684.

[39] Cha, M., Benevenuto, F., Haddadi, H., and Gummadi, K. (2012). The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(4):991--998.

[40] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 10--17.

[41] Chakraborty, A., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2016a). Dissemination biases of social media channels: On the topical coverage of socially shared news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 559–562.

[42] Chakraborty, A., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2017a). Optimizing the recency-relevancy trade-off in online news recommendations. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 837--846.

[43] Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2017b). Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 22--31.

[44] Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. (2016b). Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9--16.

[45] Chauhan, A. and Hughes, A. L. (2020). Trustworthiness perceptions of social media resources named after a crisis event. *Proceedings of the ACM on Human-Computer Interaction*, 4:1--23.

[46] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785--794.

[47] Chen, Y., Conroy, N. J., and Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as false news. In *Proc. of the ACM on Workshop on Multimodal Deception Detection (WMDD)*, pages 15--19.

[48] Cheng, A., Evans, M., and Singh, H. (2009). Inside twitter: An in-depth look inside the twitter world. [Report of Sysomos - Online] http://www. sysomos. com/inside-twitter.

[49] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):e0128193.

[50] Cohen, S., Li, C., Yang, J., and Yu, C. (2011). Computational journalism: A call to arms to database researchers. In *Proc. of the Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 148--151.

[51] Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

[52] Collins (2017). Collins english dictionary: Word of the year 2017. https://www.collinsdictionary.com/woty.

[53] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*, pages 1--4.

[54] Constantinides, M. (2015). Apps with habits: Adaptive interfaces for news apps. In *Proc. of the Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, pages 191--194.

[55] Constantinides, M., Dowell, J., Johnson, D., and Malacria, S. (2015). Habito news: A research tool to investigate mobile news reading. In *Proc. of the Int'l Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pages 598--598.

[56] Correa, D., Silva, L. A., Mondal, M., Benevenuto, F., and Gummadi, K. P. (2015). The many shades of anonymity: Characterizing anonymous social media content. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 71–80.

[57] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on Information Theory*, 13(1):21--27.

[58] Covert, T. J. A. and Wasburn, P. C. (2007). Measuring media bias: A content analysis of time and newsweek coverage of domestic social issues, 1975–2000. *Social science quarterly*, 88(3):690--706.

[59] Cui, L. and Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset.

[60] Cui, L., Shu, K., Wang, S., Lee, D., and Liu, H. (2019). defend: A system for explainable fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 2961--2964.

[61] Cunha, E., Magno, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2012). A gender based study of tagging behavior in twitter. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, pages 323--324.

[62] Cunha, E., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (2018). Fake news as we feel it: perception and conceptualization of the term" fake news" in the media. In *Proc. of the Int'l Conference on Social Informatics (SocInfo)*, pages 151--166. Springer.

[63] Dai, E., Sun, Y., and Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 853--862.

[64] Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2017). A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology*, 68(2):286--308.

[65] Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 271--280.

[66] De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., and Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness. In *Proc. of the Int'l ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 353--369.

[67] de Melo, P. O. V. (2015). How many political parties should brazil have? a data-driven method to assess and reduce fragmentation in multi-party political systems. *PloS one*, 10(10):e0140217.

[68] De Nies, T., Haesendonck, G., Godin, F., De Neve, W., Mannens, E., and Van de Walle, R. (2013). Towards automatic assessment of the social media impact of news content. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 871--874.

[69] Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Nature Scientific Reports*, 6:37825.

[70] Dewey, C. (2016). What we really see when Facebook Trending picks stories for us. washingtonpost.com/news/ the-intersect/wp/2016/05/20/what-we-really-see-when-facebook-trending-picks-stories-for-us.

[71] Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2451--2460.

[72] Diakopoulos, N., Naaman, M., and Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 115--122.

[73] Dixon, E., Enos, E., and Brodmerkle, S. (2011). A/b testing of a webpage. US Patent 7,975,000.

[74] Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J. B., and Vujnovic, M. (2008). Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers. *Journalism practice*, 2(3):326--342.

[75] Dong, H., Zhu, J., Tang, Y., Xu, C., Ding, R., and Chen, L. (2015). Ubs: a novel news recommendation system based on user behavior sequence. In *Proc. of the Int'l Conference on Knowledge Science, Engineering and Management (KSEM)*, pages 738--750.

[76] dos Reis, J. C. S., Gonçalves, P., Olmo, P., Prates, R., and Benevenuto, F. (2014). Magnet news: You choose the polarity of what you read. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 652--653.

[77] Došen, Đ. O. and Lidija, B. (2018). Key design elements of daily newspapers: Impact on the reader's perception and visual impression. *KOME- An International Journal of Pure Communication Inquiry*, 6(2):62.

[78] Ebrahimi, M., Yazdavar, A. H., and Sheth, A. (2017). Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, 32(5):70--75.

[79] Esiyok, C., Kille, B., Jain, B.-J., Hopfgartner, F., and Albayrak, S. (2014). Users' reading habits in online news portals. In *Proc. of the Information Interaction in Context Symposium (IIiX)*, pages 263--266.

[80] Facebook (2016). Search fyi: An update to trending. newsroom.fb.com/news/2016/08/search-fyi-an-update-to-trending/.

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

[82] Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *First Monday*.

[83] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96--104.

[84] Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1163--1168.

[85] Finn, S., Metaxas, P. T., Mustafaraj, E., O'Keefe, M., Tang, L., Tang, S., and Zeng, L. (2014). Trails: A system for monitoring the propagation of rumors on twitter. In *Computation and Journalism Symposium (C+J)*.

[86] Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298--320.

[87] Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

[88] Friggeri, A., Adamic, L. A., Eckles, D., and Cheng, J. (2014). Rumor cascades. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 101--110.

[89] Fry, E. (1968). A readability formula that saves time. *Journal of reading*, 11(7):513--578.

[90] Fujita, S., Kobayashi, H., and Okumura, M. (2019). Dataset creation for ranking constructive news comments. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2619--2626.

[91] Gallagher, K. (2017). The social media demographics report: Differences in age, gender, and income at the top platforms. http://www.businessinsider.com/the-social-media-demographics-report-2017-8.

[92] Gao, W., Li, P., and Darwish, K. (2012). Joint topic modeling for event summarization across news and social media streams. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 1173--1182.

[93] Garcin, F., Galle, F., and Faltings, B. (2014). Focal: A personalized mobile news reader. In *Proc. of the Int'l ACM Conference on Recommender Systems (RecSys)*, pages 369--370.

[94] Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 913--922.

[95] Garimella, K. and Tyson, G. (2018). Whatapp doc? a first look at whatsapp public group data. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 511--517.

[96] Gebremeskel, G. G. and de Vries, A. P. (2015). The role of geographic information in news consumption. In *Proc. of the Int'l ACM Conference on World Wide Web Conference (WWW)*, pages 755--760.

[97] Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35--71.

[98] Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. 1:623--645.

[99] Giełczyk, A., Wawrzyniak, R., and Choraś, M. (2019). Evaluation of the existing tools for fake news detection. In *Proc. of the IFIP Int'l Conference on Computer Information Systems and Industrial Management (CISIM)*, pages 144--151.

[100] Gilbert, E., Bakhshi, S., Chang, S., and Terveen, L. (2013). I need to try this?: a statistical overview of pinterest. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2427--2436.

[101] Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., et al. (2018). Fake news vs satire: A dataset and analysis. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 17--21.

[102] Gottfried, J. and Shearer, E. (2016). News use across social media platforms 2016. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

[103] Gruppi, M., Horne, B. D., and Adalı, S. (2020). Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles.

[104] Guimarães, S., Reis, J. C. S., Lima, L., Ribeiro, F., Vasconcelos, M., An, J., Kwak, H., and Benevenuto, F. (2020). Identifying and characterizing alternative

news media on facebook. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM).*

[105] Gunning, R. (1952). The technique of clear writing. *McGraw-Hill, New York.*

[106] Guo, L., A. Rohde, J., and Wu, H. D. (2020). Who is responsible for twitter's echo chamber problem? evidence from 2016 us election networks. *Information, Communication & Society*, 23(2):234--251.

[107] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. In *Proc. of the Int'l Conference on Social Informatics (SocInfo)*, pages 228--243. Springer.

[108] Hargreaves, E., Menasché, D., Neglia, G., and Agosti, C. (2018). Visibilidade no facebook: Modelos, medições e implicações. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BrasNam).*

[109] Hart, P. S., Chinn, S., and Soroka, S. (2020). <? covid19?> politicization and polarization in covid-19 news coverage. *Science Communication*, page 1075547020950735.

[110] Hassan, N., Adair, B., Hamilton, J., Li, C., Tremayne, M., Yang, J., and Yu, C. (2015). The quest to automate fact-checking. In *Computation and Journalism Symposium (C+J).*

[111] Hassan, N., Sultana, A., Wu, Y., Zhang, G., Li, C., Yang, J., and Yu, C. (2014). Data in, fact out: automated monitoring of facts by factwatcher. *Proc. of the VLDB Endowment*, 7(13):1557--1560.

[112] Hermida, A., Fletcher, F., Korell, D., and Logan, D. (2012). Share, like, recommend: Decoding the social media news consumer. *Journalism studies*, 13(5-6):815--824.

[113] Herrero-Diz, P., Conde-Jiménez, J., and Reyes de Cózar, S. (2020). Teens' motivations to spread fake news on whatsapp. *Social Media+ Society*, 6(3):2056305120942879.

[114] Horne, B. D. and Adali, S. (2017). This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proc. of the Worshops of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM): News and Public Opinion*, pages 759--766.

[115] Hosni, A. I. E. and Li, K. (2020). Minimizing the influence of rumors during breaking news events in online social networks. *Knowledge-Based Systems*, 193:105452.

[116] Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., and Ma, K.-L. (2012). Breaking news on twitter. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2751--2754.

[117] Huh, M., Liu, A., Owens, A., and Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 101--117.

[118] Hui, P.-M., Shao, C., Flammini, A., Menczer, F., and Ciampaglia, G. L. (2018). The hoaxy misinformation and fact-checking diffusion network. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 528--530.

[119] Jeronimo, C. L. M., Marinho, L. B., Campelo, C. E., Veloso, A., and da Costa Melo, A. S. (2019). Fake news classification based on subjective language. In *Proc. of the Int'l Conference on Information Integration and Web-based Applications & Services (WAS)*, pages 15--24.

[120] Jin, Z., Cao, J., Jiang, Y.-G., and Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *Proc. of the IEEE Int'l Conference on Data Mining (ICDM)*, pages 230--239.

[121] Jin, Z., Cao, J., Zhang, Y., Zhou, J., and Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598--608.

[122] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the European Conference on Machine Learning (ECML)*, pages 137--142.

[123] Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., and Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 53--54.

[124] Kazai, G., Yusof, I., and Clarke, D. (2016). Personalised news and blog recommendations based on user location, facebook and twitter user profiling. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1129--1132.

[125] Keegan, J. (2016). Blue Feed, Red Feed - see Liberal Facebook and Conservative Facebook. http://graphics.wsj.com/blue-feed-red-feed.

[126] Kim, J. H., Mantrach, A., Jaimes, A., and Oh, A. (2016). How to compete online for news audience: Modeling words that attract clicks. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1645--1654.

[127] Kirchner, J. and Reuter, C. (2020). Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1--27.

[128] Kothari, A., Magdy, W., Darwish, K., Mourad, A., and Taei, A. (2013). Detecting comments on news articles in microblogs. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 293--302.

[129] Kourogi, S., Fujishiro, H., Kimura, A., and Nishikawa, H. (2015). Identifying attractive news headlines for social media. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 1859--1862.

[130] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767.

[131] Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 591--602.

[132] Kwak, H. and An, J. (2014). A first look at global news coverage of disasters by using the gdelt dataset. In *Proc. of the Int'l Conference on Social Informatics (SocInfo)*, pages 300--308.

[133] Kwak, H. and An, J. (2016). Revealing the hidden patterns of news photos: Analysis of millions of news photos through gdelt and deep learning-based vision apis. In *Proc. of the Worshops of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM): News and Public Opinion*, pages 99--107.

[134] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 591--600.

[135] Kwon, S., Cha, M., and Jung, K. (2017). Rumor detection over varying time windows. *PloS One*, 12(1):e0168344.

[136] Kwon, S., Cha, M., Jung, K., Chen, W., and Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *Proc. of the IEEE Int'l Conference on Data Mining*, pages 1103--1108.

[137] Lagun, D. and Lalmas, M. (2016). Understanding and measuring user engagement and attention in online news reading. In *Proc. of the Int'l ACM Conference on Web Search and Data Mining (WSDM)*, pages 113--122.

[138] Lamichhane, K. P. and Shrestha, K. (2020). Implementation of machine learning approach to detect clickbaits in online news. *Fuse Machines Inc*, pages 15--19.

[139] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094--1096.

[140] Lee, C. S. and Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in human behavior*, 28(2):331--339.

[141] Lee, J. (2020). "friending" journalists on social media: Effects on perceived objectivity and intention to consume news. *Journalism Studies*, pages 1--17.

[142] Lehmann, J., Castillo, C., Lalmas, M., and Zuckerman, E. (2013). Transient news crowds in social media. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 351--360.

[143] Lerman, K. and Hogg, T. (2010). Using a model of social dynamics to predict popularity of news. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 621--630.

[144] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 497--506.

[145] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proc. of the Int'l Conference on World Wide Web (WWW)*, pages 661--670.

[146] Li, Q., Liu, X., Fang, R., Nourbakhsh, A., and Shah, S. (2016a). User behaviors in newsworthy rumors: A case study of twitter. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 627--630.

[147] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016b). A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1--16.

[148] Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., and Han, J. (2015). On the discovery of evolving truth. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 675--684.

[149] Li, Z., Wang, B., Li, M., and Ma, W.-Y. (2005). A probabilistic model for retrospective news event detection. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 106--113.

[150] Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., and Jiang, P. (2019). A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proc. of the Int'l ACM Conference on Recommender Systems (RecSys)*, pages 20--28.

[151] Liu, J., Dolan, P., and Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proc. of the Int'l Conference on Intelligent User Interfaces (IUI)*, pages 31--40.

[152] Liu, W. and Ruths, D. (2013). What's in a name? using first names as features for gender inference in twitter. In *Proc. of the AAAI Spring Symposium: Analyzing Microtext*, pages 10--16.

[153] Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., Pomerville, S., and Wudali, R. e. a. (2016). Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 207--216.

[154] Llewellyn, C., Grover, C., and Oberlander, J. (2014). Summarizing newspaper comments. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 599--602.

[155] Lokniti, C. (2018). How widespread is whatsapp's usage in india? https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html.

[156] Lu, Y.-J. and Li, C.-T. (2020). Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. pages 505--514.

[157] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proc. of the Neural Information Processing Systems (NIPS)*, pages 1--10.

[158] Lynch, D. (1993). Catch 22?: Washington newswomen and their news sources. *Newspaper Research Journal*, 14(3-4):82--92.

[159] Ma, L., Lee, C. S., and Goh, D. H.-L. (2011). That's news to me: the influence of perceived gratifications and personal experience on news sharing in social media. In *Proc. of the Int'l ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 141--144.

[160] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579--2605.

[161] Maros, A., Almeida, J., Benevenuto, F., and Vasconcelos, M. (2020). Analyzing the use of audio messages in whatsapp groups. In *Proc. of the ACM Web Conference (WWW)*, pages 3005--3011.

[162] Massey Jr, F. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68--78.

[163] Mc Laughlin, G. H. (1969). Smog grading-a new readability formula. *Journal of reading*, 12(8):639--646.

[164] McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. 752(1):41--48.

[165] McCreadie, R., Macdonald, C., and Ounis, I. (2013). News vertical search: when and what to display to users. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 253--262.

[166] Melo, P., Dalip, D., Junior, M., Gonçalves, M., and Benevenuto, F. (2019a). 10sent: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *Journal of the Association for Information Science and Technology (JAIST)*, 70(3):242--255.

[167] Melo, P., Messias, J., Resende, G., Garimella, K., Almeida, J., and Benevenuto, F. (2019b). Whatsapp monitor: A fact-checking system for whatsapp. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 676--677.

[168] Melo, P., Vieira, C. C., Garimella, K., de Melo, P. O., and Benevenuto, F. (2019c). Can whatsapp counter misinformation by limiting message forwarding? In *Proc. of the Int'l Conference on Complex Networks and their Applications (Complex Networks)*, pages 372--384.

[169] Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 554--557.

[170] Mitchell, A. (2016). Key findings on the traits and habits of the modern news consumer. http://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/.

[171] Mitra, T. and Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 258--267.

[172] Monteiro, R. A., Santos, R. L., Pardo, T. A., de Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Proc. of the Int'l Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 324--334.

[173] Moreno, A., Garrison, P., and Bhat, K. (2017). Whatsapp for monitoring and response during critical events: Aggie in the ghana 2016 election. In *Proc. of Int'l Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 645--655.

[174] Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased?: assessing the representativeness of twitter's streaming api. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 555--556.

[175] Mourão, R. R. and Harlow, S. (2020). Awareness, reporting, and branding: Exploring influences on brazilian journalists' social media use across platforms. *Journal of Broadcasting & Electronic Media*, pages 1--21.

[176] Munson, S. A., Lee, S. Y., and Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 419--428.

[177] Myslinski, L. J. (2012). Fact checking method and system. Google Patents. US Patent 8,185,448.

[178] Nallapati, R., Feng, A., Peng, F., and Allan, J. (2004). Event threading within news topics. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 446--453.

[179] Naseri, M. and Zamani, H. (2019). Analyzing and predicting news popularity in an instant messaging service. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1053--1056.

[180] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, number 8.

[181] Nelson, J. L. (2020). The persistence of the popular in mobile news consumption. *Digital Journalism*, 8(1):87--102.

[182] Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211--236.

[183] Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism.

[184] Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2):141--161.

[185] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175--220.

[186] Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.-Y., Kapadia, A., and Rojas, F. (2016). Twitter's glass ceiling: The effect of perceived gender on online visibility. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 289--298.

[187] Nørregaard, J., Horne, B. D., and Adalı, S. (2019). Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 630--638.

[188] Olteanu, A., Castillo, C., Diakopoulos, N., and Aberer, K. (2015). Comparing events coverage in online news and social media: The case of climate change. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 288--297.

[189] Ong, T.-H., Chen, H., Sung, W.-k., and Zhu, B. (2005). Newsmap: a knowledge map for online news. *Decision Support Systems*, 39(4):583--597.

[190] Orellana-Rodriguez, C., Greene, D., and Keane, M. T. (2016). Spreading the news: how can journalists gain more engagement for their tweets? In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 107--116.

[191] Oxford (2019). Oxford dictionaries: “clickbait”. http://www.oxforddictionaries.com/definition/english/clickbait.

[192] Oxford (2020). Oxford dictionaries: “memes”. https://en.oxforddictionaries.com/definition/meme.

[193] Palda, K. F. (2011). *Pareto's Republic and the new Science of Peace*. Filip Palda.

[194] Park, S., Kang, S., Chung, S., and Song, J. (2009). Newscube: delivering multiple aspects of news to mitigate media bias. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 443--452.

[195] Park, S., Ko, M., Lee, J., Choi, A., and Song, J. (2013). Challenges and opportunities of local journalism: a case study of the 2012 korean general election. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 286--295.

[196] Pasquini, C., Brunetta, C., Vinci, A. F., Conotter, V., and Boato, G. (2015). Towards the verification of image integrity in online news. In *Proc. of the IEEE Int'l Conference on Multimedia Expo Workshops (ICMEW)*, pages 1--6.

[197] Paul, S., Joy, J. I., Sarker, S., Ahmed, S., Das, A. K., et al. (2019). Fake news detection in social media using blockchain. In *Proc. of the Int'l IEEE Conference on Smart Computing & Communications (ICSCC)*, pages 1--5.

[198] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

[199] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

[200] Pereira, M. H. R., Pádua, F. L. C., Pereira, A. C. M., Benevenuto, F., and Dalip, D. H. (2016). Fusing audio, textual, and visual features for sentiment analysis of news videos. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 659--662.

[201] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. pages 3391--3401.

[202] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 231--240.

[203] Poynter (2020). Fighting the infodemic: The coronavirusfacts alliance. https://www.poynter.org/coronavirusfactsalliance/.

[204] Prasojo, R. E., Kacimi, M., and Nutt, W. (2015). Entity and aspect extraction for organizing news comments. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 233--242.

[205] Pratiwi, I. Y. R., Asmara, R. A., and Rahutomo, F. (2017). Study of hoax news detection using naïve bayes classifier in indonesian language. In *Proc. of the IEEE Int'l Conference on Information & Communication Technology and System (ICTS)*, pages 73--78.

[206] Provost, F. and Kohavi, R. (1998). Glossary of terms. *Journal of Machine Learning*, 30(2-3):271--274.

[207] Przybyla, P. (2020). Capturing the style of fake news. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 490--497.

[208] Qin, T., Liu, T.-Y., Xu, J., and Li, H. (2010). Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346--374.

[209] Quattrociocchi, W., Scala, A., and Sunstein, C. R. (2016). Echo chambers on facebook. https://ssrn.com/abstract=2795110.

[210] Quezada, M., Peña-Araya, V., and Poblete, B. (2015). Location-aware model for news events in social media. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 935--938.

[211] Ramachandran, A. and Feamster, N. (2006). Understanding the network-level behavior of spammers. 36(4):291--302.

[212] Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and tracking political abuse in social media. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 297--304.

[213] Reis, J., Benevenuto, F., de Melo, P. O., Prates, R., Kwak, H., and An, J. (2015). Breaking the news: First impressions matter on online news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 357--366.

[214] Report, D. N. (2018). Statistic of the week: How brazilian voters get their news. https://reutersinstitute.politics.ox.ac.uk/risj-review/statistic-week-how-brazilian-voters-get-their-news.

[215] Resende, G., Melo, P., Reis, J. C. S., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019a). Analyzing textual (mis)information shared in whatsapp groups. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 225--234.

[216] Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019b). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proc. of the ACM Web Conference (WWW)*, pages 818--828.

[217] Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 387--390.

[218] Resnick, P., Garrett, R. K., Kriplean, T., Munson, S. A., and Stroud, N. J. (2013). Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proc. of the Int'l ACM Conference on Computer Supported Cooperative Work (CSCW) Companion*, pages 95--100.

[219] Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 290--299.

[220] Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., and Benevenuto, F. (2016). Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. 5(1):1--29.

[221] Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., and Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT)*, pages 140--149.

[222] Ribeiro, M. H., Calais, P. H., Almeida, V. A., and Meira Jr, W. (2017). "everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. In *Proc. of the Workshop on Data Science + Journalism @KDD*.

[223] Ribeiro, M. T., Ziviani, N., Moura, E. S. D., Hata, I., Lacerda, A., and Veloso, A. (2014). Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1--20.

[224] Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K. P., and Almeida, V. (2011). On word-of-mouth based discovery of the web. In *Proc. of the Int'l ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*, pages 381--396.

[225] Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011). Influence and passivity in social media. In *Proc. of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 18--33.

[226] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and Applied Mathematics*, 20:53--65.

[227] Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proc. of the Workshop on Computational Approaches to Deception Detection (NAACL-HLT)*, pages 7--17.

[228] Rubin, V. L., Chen, Y., and Conroy, N. J. (2015). Deception detection for news: three types of fakes. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*, page 83. American Society for Information Science.

[229] Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proc. of the Int'l ACM Conference on Information and Knowledge Management (CIKM)*, pages 797--806.

[230] Sáenz, C. A. C., Dias, M., and Becker, K. (2020). Combining compact news representations generated using distilbert and topological features to classify fake news. In *Proc. of the Symposium on Knowledge Discovery, Mining and Learning (KDMILE)*, pages 209--216.

[231] Saez-Trumper, D. (2014). Fake tweet buster: a webtool to identify users promoting fake news on twitter. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, pages 316--317.

[232] Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). Fa-kes: A fake news dataset around the syrian war. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 573--582.

[233] Sales, A., Balby, L., and Veloso, A. (2019). Media bias characterization in brazilian presidential elections. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, pages 231--240.

[234] Samonte, M. J. C. (2018). Polarity analysis of editorial articles towards fake news detection. In *Proc. of the Int'l Conference on Internet and e-Business (EEE)*, pages 108--112.

[235] Santia, G. and Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 531--540.

[236] Saranya, K. and Sadasivam, G. S. (2017). Personalized news article recommendation with novelty using collaborative filtering based rough set theory. *Mobile Networks and Applications*, 22(4):719--729.

[237] Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, pages 1--22.

[238] Schwartz, R., Naaman, M., and Teodoro, R. (2015). Editorial algorithms: Using social media to discover and report local news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 407--415.

[239] Senter, R. and Smith, E. A. (1967). Automated readability index. Technical report, Cincinnati Univ Oh.

[240] Shah, D. V., Cappella, J. N., Neuman, W. R., Soroka, S., Young, L., and Balmas, M. (2015). Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. *The Annals of the American Academy of Political and Social Science*, 659(1):108--121.

[241] Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 745--750.

[242] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787.

[243] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1--42.

[244] Sheidin, J., Lanir, J., Kuflik, T., and Bak, P. (2017). Visualizing spatial-temporal evaluation of news stories. In *Proc. of the Int'l Conference on Intelligence User Interfaces (IUI) Companion*, pages 65--68.

[245] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019a). defend: Explainable fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 395--405.

[246] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017a). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22--36.

[247] Shu, K., Wang, S., and Liu, H. (2017b). Exploiting tri-relationship for fake news detection. In *arXiv preprint arXiv:1712.07709*.

[248] Shu, K., Wang, S., and Liu, H. (2019b). Beyond news contents: The role of social context for fake news detection. In *Proc. of the Int'l ACM Conference on Web Search and Data Mining (WSDM)*, pages 312--320.

[249] Silva, M., Santos de Oliveira, L., Andreou, A., Vaz de Melo, P. O., Goga, O., and Benevenuto, F. (2020). Facebook ads monitor: An independent auditing system for political ads on facebook. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 224--234.

[250] Silverman, C., Strapagiel, L., Shaban, H., Hall, E., , and Singer-Vine, J. (2016). Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis.

[251] Slaughter, G. (1994). Perspectives on electronic publishing: Standards, solutions, and more. *Technical Communication*, 41(2):304--306.

[252] Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545.

[253] Smadja, U., Grusky, M., Artzi, Y., and Naaman, M. (2019). Understanding reader backtracking behavior in online news articles. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 3237--3243.

[254] Spinde, T., Hamborg, F., Donnay, K., Becerra, A., and Gipp, B. (2020a). Enabling news consumers to view and understand biased news coverage: A study on the perception and visualization of media bias. In *Proc. of the Int'l ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 389--392.

[255] Spinde, T., Hamborg, F., and Gipp, B. (2020b). An integrated approach to detect media bias in german news articles. In *Proc. of the Int'l ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 505--506.

[256] Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150--160.

[257] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood)*.

[258] Tardaguila, C., Benevenuto, F., and Ortellado, P. (2018). Fake news is poisoning brazilian politics. whatsapp can stop it.

[259] Tatar, A., Antoniadis, P., De Amorim, M. D., and Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1):174.

[260] Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., and Fdida, S. (2011). Predicting the popularity of online articles based on user comments. In *Proc. of the Int'l Conference on Web Intelligence, Mining and Semantic (WIMS)*, number 67.

[261] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24--54.

[262] Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. (2008). Newsstand: A new view on news. In *Proc. of the ACM*

SIGSPATIAL International Conference on Advances in Geographic Information Systems, number 18.

[263] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544--2558.

[264] Tian, L., Zhang, X., Wang, Y., and Liu, H. (2020). Early detection of rumours on twitter via stance transfer learning. In *Proc. of the European Conference on Information Retrieval (ECIR)*, pages 575--588.

[265] Tolmie, P., Procter, R., Randall, D. W., Rouncefield, M., Burger, C., Wong Sak Hoi, G., Zubiaga, A., and Liakata, M. (2017). Supporting the use of user generated content in journalistic practice. In *Proc. of the ACM CHI Conference on Human Factors in Computing Systems*, pages 3632--3644.

[266] Torrey, L. and Shavlik, J. (2010). Transfer learning. pages 242--264.

[267] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 517--524.

[268] Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297--323.

[269] Twitter (2010). To Trend or Not to Trend. blog.twitter.com/2010/to-trend-or-not-to-trend.

[270] Väätäjä, H. and Egglestone, P. (2012). Briefing news reporting with mobile assignments: perceptions, needs and challenges. In *Proc. of the Int'l ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 485--494.

[271] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 280--289.

[272] Veloso, B. M., Assunção, R. M., Ferreira, A. A., and Ziviani, N. (2019). In search of a stochastic model for the e-news reader. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(6):1--27.

[273] Venkatadri, G., Andreou, A., Liu, Y., Mislove, A., Gummadi, K. P., Loiseau, P., and Goga, O. (2018). Privacy risks with facebook's pii-based targeting: Auditing

a data broker's advertising interface. In *Proc. of the IEEE Symposium on Security and Privacy (SP)*, pages 89--107.

[274] Vikatos, P., Messias, J., Miranda, M., and Benevenuto, F. (2017). Linguistic diversities of demographic groups in twitter. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, pages 275--284.

[275] Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proc. of the ACL Workshop on Language Technologies and Computational Social Science*, pages 18--22.

[276] Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 647--653.

[277] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146--1151.

[278] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 454--463.

[279] Wan, X. and Zhang, J. (2014). Ctsum: extracting more certain summaries for news articles. In *Proc. of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 787--796.

[280] Wang, L. X., Ramachandran, A., and Chaintreau, A. (2016). Measuring click and share dynamics on social media: a reproducible and validated approach. In *Proc. of the Worshops of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM): News and Public Opinion*, pages 108--113.

[281] Wang, P., Angarita, R., and Renna, I. (2018a). Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 1557--1561.

[282] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 422--426.

[283] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018b). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 849--857.

[284] Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., and Gao, J. (2020). Weak supervision for fake news detection via reinforcement learning. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 516--523.

[285] Ward, A., Ross, L., Reed, E., Turiel, E., and Brown, T. (1997). Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, pages 103--135.

[286] Wei, W. and Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. In *Proc. of the Int'l Joint Conference on Artificial Intelligence (IJCAI)*, pages 4172--4178.

[287] Westlund, O. (2013). Mobile news: A review and model of journalism in an age of mobile media. *Digital journalism*, 1(1):6--26.

[288] Williams, M. J., Cioroianu, I., and Williams, H. T. (2016). Different news for different views: Political news-sharing communities on social media through the uk general election in 2015. In *Proc. of the Worshops of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM): News and Public Opinion*, pages 118--125.

[289] Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW)*, pages 705--714.

[290] Wu, Y., Agarwal, P. K., Li, C., Yang, J., and Yu, C. (2014). Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589--600.

[291] Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. (2019). Xfake: explainable fake news detector with visualizations. In *Proc. of the ACM Web Conference (WWW)*, pages 3600--3604.

[292] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69--90.

[293] Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems and their Applications*, 14(4):32--43.

[294] Yin, Q., Cao, Z., Jiang, Y., and Fan, H. (2016). Learning deep face representation. US Patent 9,400,919.

[295] Yoo, S. and Harman, M. (2007). Pareto efficient multi-objective test case selection. In *Proc. of the Int'l Symposium on Software Testing and Analysis (ISSTA)*, pages 140--150.

[296] Zagheni, E., Garimella, V. R. K., Weber, I., et al. (2014). Inferring international and internal migration patterns from twitter data. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 439--444.

[297] Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

[298] Zames, G., Ajlouni, N., Ajlouni, N., Ajlouni, N., Holland, J., Hills, W., and Goldberg, D. (1981). Genetic algorithms in search, optimization and machine learning. *Information Technology Journal*, 3(1):301--302.

[299] Zauner, C. (2010). Implementation and benchmarking of perceptual image hash functions. *Master's thesis, Upper Austria University of Applied Sciences. http://phash.org/docs/pubs/thesis_zauner.pdf.*

[300] Zeldes, G. A., Fico, F., and Diddi, A. (2007). Race and gender: An analysis of the sources and reporters in local television coverage of the 2002 michigan gubernatorial campaign. *Mass Communication & Society*, 10(3):345--363.

[301] Zhang, H. and Setty, V. (2016). Finding diverse needles in a haystack of comments: social media exploration for news. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*, pages 286--290.

[302] Zhang, J., Cui, L., Fu, Y., and Gouza, F. B. (2018). Fake news detection with deep diffusive network model. In *arXiv preprint arXiv:1805.08751.*

[303] Zhao, Z., Resnick, P., and Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*, pages 1395--1405.

[304] Zhou, D. X., Resnick, P., and Mei, Q. (2011). Classifying the political leaning of news articles and users from user votes. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 417--424.

[305] Zhou, X. and Zafarani, R. (2019). Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, 21(2):48--60.

[306] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.

Appendix A

Demographics of News Sharing in Twitter

The widespread adoption and dissemination of online news through social media systems have been revolutionizing many segments of our society and ultimately our daily lives. In these systems, users can play a central role as they share content to their friends. Despite that, little is known about news spreaders in social media. In this paper, we provide the first of its kind in-depth characterization of news spreaders in social media. In particular, we investigate their demographics, what kind of content they share, and the audience they reach. Among our main findings, we show that males and white users tend to be more active in terms of sharing news, biasing the news audience to the interests of these demographic groups. Our results also quantify differences in interests of news sharing across demographics, which has implications for personalized news digests.

A.1 Introduction

In recent years, with the huge success of Twitter and Facebook, social media has become one of the most important channels in news diffusion. In particular, Twitter's unique concepts of asymmetric "follow" and "retweet", which were later adopted by Facebook, allow users to follow each other's updates and propagate interesting pieces of information quickly and broadly [134]. Such great power to disseminate information embedded in social media naturally has attracted the news media. As a result, a majority of US adults (62%) get news mostly on social media, according to a new survey by Pew Research Center [102].

Along with their traditional channels, news media manage their presence in social media by creating Twitter accounts and publishing tweets containing URLs that link their news media sites. For those accounts, it is clearly visible who the audience is – their followers. Furthermore, as any Twitter user can share URLs to news media web sites, Twitter users exposed to news media's tweets through retweets can also be visible and accounted as audience. We call these users *news spreaders* in the rest of this paper. This form of sharing of news URLs has long been a pervasive practice in social media, but its role and impact are relatively unexplored.

In this work, we characterize news spreaders in Twitter along three dimensions: (i) their demographics (who they are), (ii) their news shared (what they share), and (iii) their impact (why they are important). To this end, the inference of demographics of Twitter users is essential. Among various techniques that have been proposed [169], we use state-of-the-art techniques to locate Twitter users and infer their demographics based on profile photos.

Through a longitudinal data collection of news spreaders and their URL sharing behavior of five popular global news media, we test how similar news URL sharing is to typical URL sharing in terms of demographics of spreaders. We find a statistically significant trend that white males participate more in news URL sharing than other race-gender groups. This suggests that news spreaders have unique characteristics, which cannot be easily perceived for typical URL spreaders in Twitter. Thus, our work is essential to understand news spreaders correctly.

We then answer the above research questions. First, we examine demographics of news spreaders. By comparing the followers of news media accounts, we discover huge differences in terms of race-gender demographics. This suggests that we need to have a broader definition of the exposure of the news media on social media that are not only a set of followers [7] but also news spreaders. Second, we examine what kinds of news are shared by news spreaders. The properties of the pieces of news are defined along three dimensions: topics, author's (journalist's) gender and race, and linguistic analysis [199] of news headlines. These three dimensions have been discovered as important factors in news reading/sharing behavior [213, 240]. Finally, we answer how important news spreaders are for news media from the perspective of audience expansion: (i) about 59% of news spreaders do not follow news media accounts in Twitter; (ii) the audience brought by the spreaders is much bigger than that of the original followers of the news media; (iii) in addition to that the demographics of the spreaders and those of the followers are quite different, the followers of the spreaders are also substantially different from the followers of the news sources in terms of demographics. In other words, the spreaders play an important role in expanding the audience of news in Twitter, which would otherwise be very limited. Lastly, we find that the demographics of news spreaders are related to the popularity of news.

Our contributions are three-fold: (i) by using a combination of state-of-the-art techniques, we investigate in details aspects of the audience of news media in Twitter, which has been considered as in-house data so far; (ii) we suggest a robust statistical framework to test the news URL sharing behavior by comparing it with typical URL sharing behavior; and (iii) our findings show that news media should understand spreaders and their followers to capture the complete picture of their presence in news media. News media's direct followers are only the tip of the iceberg of their audience in Twitter in terms of volume and demographics.

The rest of the paper is organized as follows. Section A.2 briefly surveys related efforts. Then, we present our experiment methodology and the data gathered. The next three sections cover our results. We conclude the paper by discussing implications from our findings as well as presenting directions for future work.

A.2 Related Work

In this Section, we review existing work related to news sharing along two main dimensions.

A.2.1 News Sharing and Propagation

Social media services have made personal contacts and relationships more visible and quantifiable than ever before. Users interact by following each others' updates and passing along interesting pieces of information to their friends. This kind of word-of-mouth propagation occurs whenever a user forwards a piece of information to her friends, making users a key element in this process. Not surprisingly, a number of efforts have attempted to quantify and characterize information spread in social networks as well as the role users play in such propagation [39, 40, 224, 225, 289]. For example, Rodrigues et al. [224] showed that retweets are responsible for increasing the audience of URLs by about 2 orders of magnitude. As social media became an important channel in news diffusion, some recent research efforts attempted to investigate how news are shared in these systems. Next, we detail a few approaches that provides news sharing and propagation.

Naveed et al. [180] showed that bad news tends to spread faster in systems like Twitter. In this same year, also with the use of this same social media, Armstrong et al. [15] analyzed how online media companies employ men and women in Twitter feeds and how it connects to portrayals in news. In particular, the authors looked at how mentions of men and women on Twitter may influence mentions in news stories (e.g., newspaper, television). Through the content analysis of newspaper and television tweets at different granularity (i.e., local, regional and national), they found that male mentions were more likely to appear in national news than in regional or local news and more often than female mentions in the print media than on television.

A recent effort [32] has tackled the question "Why are some news articles shared more than others?". They showed that story importance cues are relevant in driving social sharing and that certain topics (i.e., stories about politics, accidents, disasters, and crime) were less shared. Some topics can be shared in order to improve the users' reputation. This dynamic media attention has inspired other recent studies [10]. Bright et al. [32], compare different social networks platforms and showed that some kind of news are shared more in one network than the others (e.g., economy news on LinkedIn).

Unlike previous works, our effort focuses on understanding the dynamics of news sharing on Twitter of each demographic group. Thus, to the best of our knowledge, this is the first effort that investigates intersection between news sharing and demographic information of users, including how these aspects are related.

A.2.2 Demographics in Social Media

Mislove et al. [169] was one of the first researchers that analyze demographic characteristics of Twitter users considering a geographical perspective (i.e., how the demographics vary across different US states). After that, several efforts have arisen that investigate demographic information, in various social media, using different strategies for distinct purposes [29, 34, 123]. Particularly, researchers are jointly applying computer science and statistical techniques to support sociological studies using large-scale social media datasets. These studies can range from a simple characterization of to the investigation of more complex causes, including to raising attention to the different levels at which gender biases can manifest themselves on the web [278].

In [61] the authors used Twitter data to analyze the difference between men and women behavior in terms of dynamics in free tagging environments. The results obtained present gender distinctions in the use of Twitter hashtags, emphasizing it as a social factor influencing the user's choice of specific hashtags on a specific topic. Still about tags (or hashtags), recently, the work presented in [12] explored their use by different demographic groups. The demographic characteristics of each user were obtained using *Face++* and the Twitter user's profile picture. The results showed that, although there are more popular hashtags that are commonly used, there are also many group-specific hashtags with non-negligible popularity. Besides that, the researchers show that the strategy of getting demographic data from *Face++* is reliable and provides accurate demographic information for gender and race, encouraging the application of this strategy in other recent efforts [43]. We use a similar strategy to gather demographic information.

Nilizadeh et al. [186] explore gender inequalities in Twitter, showing that gender may allow inequality to persist in terms of online visibility. Looking at Pinterest, Gilbert et al. [100] investigated what role gender plays in the website's social connections. The results highlight a major difference between female and male users regarding their motivations for using this social media. They found that being female means more repins (i.e., more shared content), but fewer followers in comparison with Twitter. Gender differences has also been explored in terms of social media disclosures of mental illness [66].

More recently, An et al. [9] examined the news consumption in South Korea (from Daum News portal[1]). The authors analyzed on a large scale the differences in news consumption from a demographic perspective. Through a multidimensional analysis of gender and age differences in news consumption, they quantify such differences along four distinct dimensions: actual news items, topic, issue, and angle. The top 30 news items for each gender and age group in Daum News were used and the demographics information were obtained through the website itself. Overall, focus mainly on quantifying and explaining differences in news consumption.

More broadly, most of the previous efforts attempt to quantify differences in gender behavior and inequalities in different social media or news systems. Our effort is the first of its kind to provide a characterization of news sharing across different demographic groups. Thus our effort is complementary to the existing ones.

A.3 Methodology

In order to understand demographics of news sharing in Twitter, first we define our strategy for data collection. Then, we define our strategies for inference of demographic information of each individual Twitter user and collection of information such as category and authors of the news, and followers of each of the news media on Twitter. Our ultimate goal, in this section, consists of reporting our baseline for comparison in order to verify the statistical significance of the results. Next, we briefly describe the methodology adopted for this work, including a discussion of its main limitations.

[1] https://media.daum.net/

A.3.1 Data Collection

For this work, we gathered the 1% random sample of all tweets, through the Twitter Streaming API[2], along a 3 months period, from July to September, 2016. Specifically, we considered only tweets (and retweets) that contain at least one URL and have been shared by U.S. users. We understand that users who share URLs may present a slight difference in behavior compared to others, so, considering our research objective, we only select this set of American users. Besides that, as we are interested in analyzing demographic characteristics, it is important to study users from the same place. For this reason, we consider only US users, filtered by timezone. In total, we retrieved 11,790,679 tweets posted by 11,770,273 US users. From this initial dataset, we infer demographics information about users and build: (i) our news sharing dataset, used in the execution of our experiments, and (ii) our baseline dataset.

A.3.2 Inferring Demographics Information

In the literature, several studies present strategies for inference of gender, race, and age. Some efforts attempt to infer the gender of a user from her name [123, 152, 169], or the age from Twitter profile descriptions [252], by using patterns like '*like 25 yr old*' or '*born in 1990*'. However, in some cases the number of unrealized inferences (e.g., for lack of information) is high (Liu and Ruths [152] reported 66% users in their dataset did not have a proper name).

To overcome such limitation, in this work, we use the profile picture's URLs of all users in our dataset and use the *Face++* API[3], a face recognition platform based on deep learning [294], to infer the gender (i.e., male or female), race (limited to Asian, Black[4], and White) and age information from the recognized faces in the profile images. We discarded users whose profile pictures do not have a recognizable face or have more than one face, according to *Face++*. Our final dataset contains 937,308 unique users located in U.S. with identified demographic information, which are gender, race, and age by *Face++*.

[2] `https://dev.twitter.com/streaming/public`

[3] `https://www.faceplusplus.com`

[4] We called *African-American (AF-AM)* in the rest of this chapter.

Table A.1: Data collection by news source.

News Media	**#Shares**	**#Authors**	**Screen name**	**#Followers**
New York Times	14,505	1,165	@nytimes	1,141
Reuters	4,712	485	@Reuters	1,259
The Guardian	4,457	844	@guardian	1,620
Wall Street Journal	1,379	313	@WSJ	1,445
BBC News	1,144	190	@BBCBreaking	1,130

Table A.2: Demographic distribution of news spreaders.

Race (%)	**Gender (%)**		**Total:**
	Male	**Female**	
Asian	5.29	6.05	11.34
AF-AM	6.09	3.80	9.89
White	43.46	35.31	78.77
Total:	54.84	45.16	100.00

A.3.3 Shared News Dataset

To focus on news sharing in Twitter, we filtered only tweets that shared news URLs from important and different news sources (i.e., BBC News[5], The New York Times[6], Reuters Online[7], The Wall Street Journal[8], The Guardian[9] and BBC News[10]), known worldwide. All these news sites appear among the most popular ones in the world, according to Alexa.com[11]. Simultaneously, we gathered information from users who posted each of the tweets including demographic information from *Face++*, as detailed above. From news URLs, we crawled information about them including, title, text, principal image (link - when there is one), authors (when there is one) and date. Lastly, Table A.1 shows the dataset used in this work containing 26,211 unique news articles shared by 16,382 unique users. We note that The New York Times is the most widely shared news media in Twitter, in comparison with those news sites considered. Table A.2 shows the demographic decomposition of those 16,382 users who shared news URLs.

[5] http://www.bbc.com

[6] http://www.nytimes.com

[7] http://www.reuters.com

[8] http://www.wsj.com/

[9] https://www.theguardian.com

[10] http://www.bbc.com

[11] http://www.alexa.com/topsites/category/News

A.3.4 Inferring News Category

In order to infer the categories of the news articles, we use meta information embedded in the news URLs. News media usually have several news sections, such as Politics, Sports, or World News, and group their news articles by these sections. By looking at which section a news article belongs to, we can infer a topical category of the news articles. The section information is often embedded in news URL. For example, the URL `http://www.nytimes.com/2016/07/02/us/politics/loretta-lynch-hillary-clinton-email-server.html` represents that the news article is about “Politics". We parsed all News URLs and extracted the topic information. The New York Times, The Guardian, and BBC adopt the above mentioned strategy for their URLs, and thus we simply parse their URL and infer the topic of a given news article. Reuters and The Wall Street Journal do not have category information in their URLs, however, the news articles have the category information. Thus, we collected news articles and extracted category information by parsing HTML files. We successfully inferred the topical categories of 93.3% (24,466) of news articles. Figure A.1 shows the proportion of the top 10 most significant news categories. We find that “World" is the most “shared” category (21.16%), similar to the results in [213].

A.3.5 Finding Journalists in Twitter

We aim to collect demographics of the authors of news articles in our dataset. Figure A.2 shows the procedure for creating an author dataset. For each news URL, we collect its title, text, principal image, authors, and date by parsing the original web page.

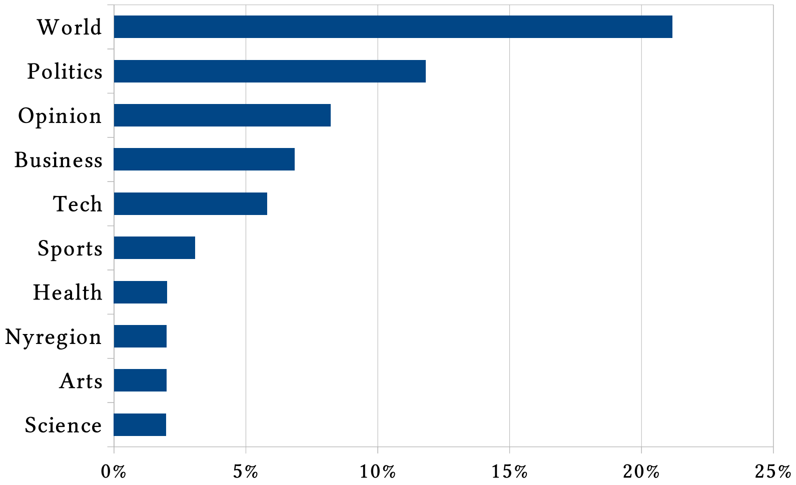


Figure A.1: Top 10 most significant news categories.

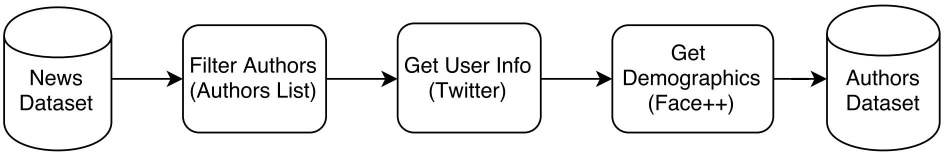


Figure A.2: Strategy for collecting news authors.

Then, we search and collect the Twitter profiles of the authors if they have Twitter accounts. Then, we infer those authors' demographic characteristics using *Face++* (see Section A.3.2). Table A.1 shows the number of authors for each news media. As expected, the largest number of names of distinguished authors we have gathered are from the The New York Times news media, which had the largest number of news shared in Twitter in our dataset.

A.3.6 Collecting Followers of News Media in Twitter

For each news source, we collected their followers in Twitter. Again, we infer their demographics by *Face++*. Table A.1 presents the total of gathered news media followers in Twitter, including the screen name used for collection. On average, we retrieved 1,319 followers by news source.

A.3.7 Baseline Dataset

A null model is widely used to estimate the statistical significance of the observed trend in given data. As the null model is randomly generated data that preserve some properties of the original data (e.g., the degree distribution in complex networks), the same trend observed from the null model captures its occurrence by chance. Then, by comparing the trend in the original data with that in the null model, the statistical significance of the observed trend in the original data can be measured.

Table A.3 shows the breakdown of ethnicity and gender of the ≈ 1 million users who shared URLs in Twitter between July and September 2016. We present a detailed description of the comparison with null models.

Table A.3: Demographic distribution of users in the Baseline dataset.

Race (%)	**Gender (%)**		**Total:**
	Male	**Female**	
Asian	7.07	10.33	17.40
AF-AM	8.52	6.93	15.45
White	31.97	35.18	67.15
Total:	47.56	52.44	100.00

In this work, whenever we report the number of users with certain properties who share URLs on particular news media, we report Z-score by comparing the number of those users in the actual data with that in null models.

Consider that we are interested in users who are Asian and share BBC News. In this case, we denote by $|U_{BBC}|$ the number of users who share BBC News and $|U_{BBC}^{Asian}|$ by the number of Asian among them. To construct a null model, we create k random samples from a separate huge set of users, which is called Population, where each sample has exactly $|U_{BBC}|$ users. The demographic information of users in Population is inferred by *Face++*. For each sample, we count how many Asians are included, $|S_{BBC}^{Asian}|$. Then, the Z_{BBC}^{Asian} is computed as following:

$$Z_{BBC}^{Asian} = \frac{|U_{BBC}^{Asian}| - mean(|S_{BBC}^{Asian}|)}{std(|S_{BBC}^{Asian}|)} \tag{A.1}$$

where $mean(\cdot)$ is the mean and $std(\cdot)$ is the standard deviation of the values from multiple samples. Intuitively, when the absolute value of Z value becomes bigger (either positive or negative), the trend (more number or less number, respectively) is less likely observed by chance. In this work, the size of Population is $\approx$ 1 million, and k=100.

A.3.8 Potential Limitations

There are a few limitations of our data, discussed next.

Accuracy of the inference by *Face++*. First, (i) we are limited by accuracy of *Face++* in the inference. *Face++* itself returns confidence levels for the inferred gender and race attributes and returns an error range for inferred age. In our data, the average confidence level reported by *Face++* is 95.24 ± 0.020% for gender and 86.12 ± 0.032% for race, with a confidence interval of 95%. Besides that, as the performance of deep learning systems continues to improve, the inferred demographic attributes should become more accurate. Also, recent efforts have used *Face++* for similar tasks and reported high confidence in manual inspections of small samples [12, 296]; Another limitation, is that (ii) *Face++* reports race of recognizable faces from images but not the *ethnicity* (e.g., *Hispanic*); Finally, though (iii) we had discarded about 70% of the crawled users (i.e., those users whose profile pictures do not have a recognizable face or have profile pictures in which *Face++* recognized with low confidence). However, we note that the remaining final dataset is still representative and we only provide results that are statistically significant based on well known statistical tests.

Data. (iv) Our approach to identify users in US may contain users located in the same time zone, but not in the US. We, however, believe that these users represent a small fraction of the users, given the predominance of active US users in Twitter [48]; *(v)* We are using the 1% random sample off all tweets. Although the 1% random sample is not the best data to capture all the dynamics happening in Twitter, its limitations are known [174] and it is the best available option at our disposal.

Even with limitations, we believe that our dataset and methods can provide interesting insights on demographics and news sharing behaviors. In the following sections, we present and discuss the main results from characterizing news spreaders in Twitter along three dimensions: (i) their demographics (who they are), (ii) their news shared (what they share), and 3) their impact (why they are important).

A.4 Who are the news spreaders?

Our first research question is to understand who the spreaders are. We compare the demographics of news spreaders with 1) the spreaders of typical URLs in Twitter and 2) the Twitter followers of news media to see whether and to what extent they differ.

A.4.1 Typical URL Sharing Vs. News Sharing

Table A.4 shows, for each news media, the proportion of news URL shares by different demographic groups. For example, for The New York Times, 54.1% of news shares are made by men and 79.2% of news shares are by Whites. The numbers in the parenthesis correspond to the Z-values, detailed in Section A.3.7. We note that the Z-value indicates how news URL sharing behavior is similar or dissimilar from typical URL sharing behavior in terms of demographic composition.

By comparing between the news sources, we see some obvious patterns: (i) The Wall Street Journal is favored by Male (62.3%) more than Female (37.7%); (ii) The New York Times has the most balanced gender distribution among spreaders (54.1% vs 45.9%); and (iii) for The New York Times, The Guardian, and BBC News, the proportion of shares by Asians is greater than by AF-AM.

From a simple comparison to Table A.2 which shows the demographic compositions of typical URL sharing behavior, we observed the following trends for all five news sources. First, Males share more news URLs than Female do. Male (54.84% of news spreaders) issue 54.1% to 62.3% of news URL shares. Secondly, Whites share

Table A.4: Proportion of news shares by different demographic groups for each news source.

News Media	**Race (%)**	**Gender (%)**		**Total:**
		Male	**Female**	
The New York Times	Asian	5.1 (-9.22)	5.9 (-18.02)	11.0 (-19.96)
	AF-AM	6.1 (-13.95)	3.7 (-15.01)	9.8 (-21.75)
	White	42.8 (26.24)	36.4 (2.86)	79.2 (31.32)
	Total:	54.1 (15.30)	45.9 (-15.30)	100.0
Reuters	Asian	3.6 (-8.06)	6.8 (-7.62)	10.4 (-12.09)
	AF-AM	7.3 (-3.02)	3.7 (-8.70)	10.9 (-9.03)
	White	47.0 (23.21)	31.7 (-4.89)	78.7 (16.38)
	Total:	57.9 (14.00)	42.1 (-14.00)	100.0
The Guardian	Asian	4.9 (-6.11)	5.9 (-9.75)	10.7 (-12.75)
	AF-AM	5.5 (-7.63)	3.3 (-9.77)	8.8 (-12.11)
	White	46.9 (23.03)	33.6 (-2.39)	80.5 (18.41)
	Total:	57.2 (13.24)	42.8 (-13.24)	100.0
The Wall Street Journal	Asian	4.9 (-3.91)	3.6 (-8.60)	8.5 (-9.43)
	AF-AM	6.1 (-3.41)	3.3 (-5.86)	9.4 (-6.68)
	White	51.3 (15.70)	30.8 (-3.35)	82.2 (12.23)
	Total:	62.3 (10.77)	37.7 (-10.77)	100.0
BBC News	Asian	5.3 (-2.64)	6.6 (-4.49)	12.0 (-5.11)
	AF-AM	7.1 (-1.91)	2.7 (-6.01)	9.8 (-5.76)
	White	46.2 (11.00)	32.1 (-2.36)	78.2 (8.04)
	Total:	58.6 (7.97)	41.4 (-7.97)	100.0

more news URLs than other race groups–White (78.77% of total users) cover 78.2% to 82.2% of news URL shares.

The Z-values in Table A.4 tell whether the differences between news spreaders and typical URL spreaders are statistically significant or not. The most strong tendency is observed for White-Male. White-Male share more news URLs than they share typical URLs and this tendency is strong (Z > 11[12]). Then, another observations is that White-Female are less likely to share news URLs than typical URLs (Z < 0) except for The New York Times. On average, White-Male make 46.8% and White-Female make 32.9% of news URL shares. From the two proportions, one may think this is because White-Female are less active than White-Male in Twitter. However, our method of comparing the news URL sharing behavior with typical URL sharing behavior can effectively tell that the difference is not because of the activity level, but of the type of URLs. White-Female do share a significant number of typical URLs.

[12] Z-value is minimum for BBC News, the largest Z-value is 26.24 for The New York Times.

Table A.5: Proportion of distinct followers by different demographic groups for each news source.

News Media	**Race (%)**	**Gender (%)**		**Total:**
		Male	**Female**	
The New York Times	Asian	12.7 (6.69)	10.5 (0.28)	23.2 (5.28)
	AF-AM	11.4 (3.71)	3.9 (-4.35)	15.2 (-0.36)
	White	35.0 (2.41)	26.6 (-6.12)	61.5 (-4.08)
	Total:	59.1 (7.97)	40.9 (-7.97)	100.0
Reuters	Asian	11.3 (5.83)	7.9 (-2.97)	19.2 (1.71)
	AF-AM	16.9 (9.97)	3.6 (-4.64)	20.5 (3.98)
	White	39.5 (5.74)	20.8 (-10.31)	60.3 (-4.52)
	Total:	67.7 (15.81)	32.3 (-15.81)	100.0
The Guardian	Asian	8.5 (2.22)	7.8 (-3.34)	16.4 (-1.30)
	AF-AM	10.5 (2.79)	3.8 (-4.58)	14.3 (-1.04)
	White	41.4 (8.99)	27.9 (-5.82)	69.3 (1.80)
	Total:	60.4 (10.45)	39.6 (-10.45)	100.0
The Wall Street Journal	Asian	9.9 (4.13)	8.0 (-3.20)	17.9 (0.54)
	AF-AM	14.5 (8.55)	4.2 (-4.06)	18.8 (3.64)
	White	41.6 (6.97)	21.7 (-11.70)	63.3 (-3.28)
	Total:	66.0 (13.93)	34.0 (-13.93)	100.0
BBC News	Asian	12.5 (5.85)	11.3 (0.92)	23.8 (4.67)
	AF-AM	12.5 (4.58)	2.2 (-6.30)	14.7 (-0.59)
	White	34.6 (1.92)	26.9 (-5.13)	61.5 (-3.25)
	Total:	59.6 (7.57)	40.4 (-7.57)	100.0

A.4.2 Are Spreaders Similar to Followers of Media Sources?

In the previous analysis, we observed that White-Male are dominant in sharing news URLs. Then, would such pattern find for the Twitter followers of news sources?

Table A.5 presents the demographics of Twitter followers of each news source. Again, the number in the parenthesis is Z-value, reporting how it differs from typical news sharing behavior. Compared to those users who share typical URLs, we observe two main differences of news media followers: 1) there are more male users ($Z > 0$); 2) except The Guardian, all the other four news sources have fewer White users ($Z < 0$). The New York Times and BBC News have more Asian followers and Reuters and The Wall Street Journal have more Asian and AF-AM users. This results in that the following three groups, Asian-Male, AF-AM-Male, and White-Male, are prominent in the followers of media sources ($Z > 0$). In addition, we observe that two news sources, The New York Times and BBC News, have positive Z-values for Asian Female followers.

For both type of users the followers and the spreaders we observe a "Male dominant" pattern, confirming that Male are more interested in news for consumption and spread. However, we find significant differences in demographic compositions between the followers and the spreaders of news. While the followers have a certain degree of racial equality, the spreaders are biased towards one particular race, White. This result is particularly important because so far it was known that individuals affiliated with news media play a large part in breaking the news [116]. Our observation indicates that breaking news is from not only those followers, but also from these news spreaders who are not necessarily following the news sources in Twitter.

A.5 What do News Spreaders Share?

We study what news spreaders share along three distinct dimensions: the topical category of news, the demographic trait of the authors (journalist) of a news article, and the linguistic properties of news headlines.

A.5.1 By News Category

We firstly examine which categories of news are shared more by particular demographic groups. To this end, we standardized the names of topical categories for the analysis. For example, we grouped news categories relating to health and life and named “Health and Life" and grouped news categories relating to science and named “Science and Tech.".

Table A.6 shows the proportion of news shares by each demographic group for each topical category. We consider only topics that were present in all news sources for this analysis. Foremost in Science and Tech, Business, and Politics, we can see the great gender differences. On average, 61.2% of news URLs of these three topics are shared by Male. In the others two categories, World and Health and Life, Female make more contributions (48.6% of shares).

When compared to typical URL sharing behavior, we observe the tendency of White-Male sharing news URLs for all categories ($Z > 0$), but the tendency is stronger for Science and Tech, Business, and Politics ($Z > 9.76$) than World ($Z = 4.55$) and Health and Life ($Z = 1.89$). One interesting observation is that White-Female do share more news URLs of World and Health and Life categories than the typical URLs ($Z > 0$).

To understand better how demographic traits relate to topical preferences, we compute the relative preferences of each demographic group to ten topical categories (see Figure A.1).

Table A.6: Number of shares by category.

Category	**Race (%)**	**Gender (%)**		**Total:**
		Male	**Female**	
World	Asian	4.3 (-6.96)	7.8 (-6.32)	12.1 (-9.84)
	AF-AM	6.1 (-6.47)	3.2 (-9.35)	9.3 (-12.53)
	White	40.4 (13.71)	38.3 (4.62)	78.6 (17.67)
	Total:	50.8 (4.55)	49.2 (-4.55)	100.0
Health and Life	Asian	6.8 (-0.18)	7.0 (-2.76)	13.8 (-2.25)
	AF-AM	3.3 (-3.82)	3.7 (-2.94)	7.0 (-5.54)
	White	41.9 (4.77)	37.3 (0.93)	79.2 (5.97)
	Total:	52.0 (1.89)	48.0 (-1.89)	100.0
Science and Tech	Asian	5.2 (-3.55)	5.4 (-6.98)	10.5 (-7.61)
	AF-AM	6.3 (-3.25)	1.7 (-10.12)	8.0 (-9.17)
	White	52.6 (19.17)	28.8 (-5.95)	81.4 (12.34)
	Total:	64.1 (15.74)	35.9 (-15.74)	100.0
Business	Asian	4.0 (-4.95)	5.3 (-6.93)	9.3 (-8.13)
	AF-AM	7.0 (-2.54)	3.1 (-5.69)	10.0 (-6.30)
	White	49.4 (15.59)	31.3 (-3.48)	80.7 (10.45)
	Total:	60.3 (9.76)	39.7 (-9.76)	100.0
Politics	Asian	5.5 (-3.06)	4.6 (-9.52)	10.1 (-9.94)
	AF-AM	6.3 (-4.01)	3.2 (-7.53)	9.5 (-8.18)
	White	47.3 (16.91)	33.1 (-2.58)	80.4 (14.20)
	Total:	59.1 (13.23)	40.9 (-13.23)	100.0

News articles about Tech are more likely to be shared by Male than Female. We then see White are more likely to share news about Health and Tech while Asian and AF-AM participate more in sharing news about Sports and Arts. Lastly, Science is favored by Asian but Business is favored by AF-AM. Our analysis shows that demographic groups have different topical tastes in sharing. This guides us how news media publish their contents to target appropriate user segments.

A.5.2 By Author's Demographics

In this section, we study how the gender of a journalist who wrote a news article influences its shares. While some differences in topics written [158] or sources used [300] between male and female journalists have been reported [158] , its appealing to each demographic group has not been fully explored.

Table A.7 shows the demographics of the authors for each news source. Overall, the proportion of Male authors are higher than that of Female authors–on average, 60.04% of the authors are Male. Reuters and BBC News have more skewed gender distributions than the other three sources. In terms of race, most of the authors are

Table A.7: Demographic characteristics of the collected authors by news source.

News Media	**Race (%)**	**Gender (%)**		**Total:**
		Male	**Female**	
The New York Times	Asian	4.9	5.8	10.7
	AF-AM	3.9	0.9	4.8
	White	49.4	35.1	84.5
	Total:	58.1	41.9	100.0
Reuters	Asian	6.8	6.0	12.8
	AF-AM	4.3	2.3	6.6
	White	51.3	29.3	80.6
	Total:	62.5	37.5	100.0
The Guardian	Asian	3.4	4.6	8.1
	AF-AM	3.8	1.2	5.0
	White	50.4	36.6	87.0
	Total:	57.6	42.4	100.0
The Wall Street Journal	Asian	7.0	6.1	13.1
	AF-AM	2.9	1.6	4.5
	White	47.9	34.5	82.4
	Total:	57.8	42.2	100.0
BBC News	Asian	3.2	4.7	7.9
	AF-AM	6.3	1.1	7.4
	White	54.7	30.0	84.7
	Total:	64.2	35.8	100.0

White (83.8% on average across five media sources), followed by Asian authors (10.5%). We observe only 5.7% of the authors are AF-AM and strikingly low fraction of AF-AM Female authors (1.42%).

Table A.8 shows the proportion of the spreaders who shared any news URLs written by a certain author demographic group for each news source.

A.5.2.1 Author's Gender

Does the gender of an author affect the spreading behavior? For The New York Times and Reuters, the proportion of Male spreaders is not significantly different ($< 2\%$) no matter the gender of the author is. However, in the rest three others sources, Male tend to share more news URLs written by Male–the difference is 12.4% for BBC News, 7.4% for The Wall Street Journal, and 5.3% for The Guardian. While the effect of the gender of the authors on spreading behavior exists, this might be a mere effect of biological differences in topical tastes–Male and Female journalists write only the topics that readers of the same gender are interested in.

To control the effect of the topics, we use a Chi-square test [37] to find which

Table A.8: Confusion matrixes for news authors and spreaders by news source.

(a) The New York Times

		Spreaders (%)		
		Male	Female	
Authors (%)	Male	54.8	45.2	
	Female	53.1	46.9	
		Asian	AF-AM	White
	Asian	11.0	10.7	78.3
	AF-AM	11.3	11.6	77.1
	White	10.7	10.3	79.1

(b) Reuters

		Spreaders (%)		
		Male	Female	
Authors (%)	Male	58.7	41.3	
	Female	57.5	42.5	
		Asian	AF-AM	White
	Asian	13.2	10.4	76.5
	AF-AM	14.0	9.6	76.3
	White	9.5	10.9	79.6

(c) The Guardian

		Spreaders (%)		
		Male	Female	
Authors (%)	Male	59.7	40.3	
	Female	54.4	45.6	
		Asian	AF-AM	White
	Asian	13.2	10.0	76.8
	AF-AM	14.1	11.1	74.8
	White	10.5	9.7	79.8

(d) The Wall Street Journal

		Spreaders (%)		
		Male	Female	
Authors (%)	Male	66.9	33.1	
	Female	59.6	40.4	
		Asian	AF-AM	White
	Asian	12.5	9.2	78.4
	AF-AM	12.3	9.9	77.8
	White	9.3	9.8	80.9

(e) BBC News

		Spreaders (%)		
		Male	Female	
Authors (%)	Male	62.4	37.6	
	Female	50.0	50.0	
		Asian	AF-AM	White
	Asian	3.4	6.9	89.7
	AF-AM	13.3	40.0	46.7
	White	11.3	9.1	79.6

Table A.9: Discriminative topics for gender and race groups by authors and spreaders.

(a) Gender

Topic	**Author**		**Spreader**	
	Female	Male	Female	Male
Sport	↓	↑		
Opinion	↓	↑		
Health	↑	↓	↑	↓
Tech			↓	↑
Business			↓	↑

(b) Race

Topic	**Author**			**Spreader**		
	Asian	AF-AM	White	Asian	AF-AM	White
World	↑		↓			
Tech	↑		↓			
Opinion	↓		↑			
Sports				↑	↑	↓
Art				↑		↓

topics are written significantly more by Female (or Male) journalists and which topics are significantly more shared by Female (or Male) spreaders. Table A.9 shows the graphical presentation of the statistically significant results by Chi-square test statistics ($p < 0.05$). In the table, an upward pointing arrow represents a higher tendency in writing or sharing. For example, Male authors write news significantly more about Sport and Opinion, and Female authors write about Health. There are no topics that authors and spreaders have the same gender differences except for Health. Therefore, the gender difference in spreading behavior is unlikely driven by that in journalists' choice of the topics. We bring the potential explanation in later section based on linguistic component of news.

A.5.2.2 Author's Race

Does the race of an author affect the spreading behavior? We observe that the proportion of Asian spreaders are significantly difference across different race of the authors in all news sources except The New York Times. For Reuters, The Guardian, and The Wall Street Journal, Asian spreaders are more likely to share news URLs written by Asian or AF-AM authors. Compared to the proportion of shares by Asian (Table A.4) which are 10.4%, 10.7%, and 8.5% for those three news sources, respectively, the proportion of the news URLs shares written by Asian authors are increased by 26.9%, 23.4%, and 47.1%, respectively. For AF-AM users, we did not find the same pattern. Lastly, BBC News has a strong tendency that AF-AM share extensively news URLs written by AF-AM and Asian.

Table A.9(b) shows the discriminative topics for each racial group of authors and spreaders. Asian authors are writing more about World and Tech than White. White authors write more opinionated news articles than Asian. For spreaders, Asian and AF-AM share more Sports news than White. News about Arts is favored by Asian more than White. Once again, we do not find any relationship between the topical interests of a certain racial author group and those of a certain racial spreaders group.

A.5.3 By LIWC Analysis

Linguistic Inquiry and Word Count (LIWC) [199] is a dictionary-based text mining software. Since it has been proposed, it has been widely used for a number of different tasks, including sentiment analysis [220] and discourse characterization in social media platforms [56]. Next, we use LIWC to characterize differences in the content shared by different demographic groups. Its latest version, LIWC 2015 (used in this work), defines about 90 linguistic categories and classifies more than 6,400 words into those categories [198]. For example, the word ‘cried’ falls into the sadness, negative emotion, overall affect, verbs, and past focus categories. Then, in a given text, the LIWC software finds the occurrence of the words in each category. The output is the proportion of the words in each category to the total words in the text.

Table A.10 presents the result of LIWC analysis of headlines shared by Male

Table A.10: LIWC analysis of ours and Newman et al. [182].

LIWC Dimension	**Our data**	**Newman et al. [182]**
Pronouns		
First-person singular	M<F	M<F
Third-person	M<F	M<F
Linguistic dimensions		
Negations	M<F	M<F
Current concerns		
Money	M>F	M>F
Biological process		
Ingestion	M<F	-
Spoken categories		
Assent	M<F	-
Swear words	M>F	M>F
Female references	M<F	-

spreaders and Female spreaders. For comparison purposes, we also show the result of effects of gender on language use [182]. We show only LIWC dimensions that have more than 20% differences between Male and Female and omit the rest because the number of the whole dimensions is more than 90.

In our data, we find exactly the same trend as [182]: Female share headlines including more first-person singular pronouns, third-person pronouns, negations, words about ingestion (e.g., dish, eat, or pizza), assent (e.g., agree, yes, or ok), and female references (e.g., girl, her, or mom), and Male share headlines including more words about money (e.g., audit, cash, or owe), and swear words (e.g., damn, or shit). Considering that [182] observed those language usage patterns in the texts Male or Female *write*, finding the same patterns in the texts he or she *shares* is surprising and interesting. The spreaders are likely to share the news that is aligned with the language usage of their own. While many research have focused on attracting more clicks by tweaking headlines, such as including named-entities in headlines [126], we show that those studies can be extended to target specific user segments.

In addition, we find some results that are aligned with some stereotypes of races (e.g., Asian share headlines including more words related to family). However, we omit the result of LIWC by race of the spreaders because there have been no available references for a systematic comparison.

A.6 Importance of Spreaders

Finally, we study the impact of understanding news spreaders in two ways: (i) extended readership by news spreaders and (ii) understanding news popularity and demographics of news spreaders.

For the first, we compare the original followers and followers of spreaders by the number and the demographics. That is, we analyzed how spreaders extend news media's readers. For example, if followers of the The New York Times are usually white male but spreaders of The New York Times URLs have a lot of Asian followers, then, the role of spreaders is really important not only because it increases the number of audience but also because it brings “different" audiences. The results are shown below in detail.

A.6.1 Extended Readership by Spreaders

Ideally, to study the audience size reached by spreaders that is not reached directly by news sources profiles, we would like to have at our disposal the followers and friends of

Table A.11: Total/Real number of followers of the news sources in Twitter and number of followers of the spreaders that shared news of the news source.

News Media	**#Followers (news media)**	**# Followers (spreaders)**
The New York Times	32,626,611	67,458,732
Reuters	15,946,449	11,119,453
The Guardian	6,154,465	21,120,210
The Wall Street Journal	12,563,525	6,193,775
BBC News	27,871,624	4,713,614

all users from our dataset. However, the number of followers and friends of these users surpasses a billion users, which is unfeasible to be crawled given our resources. As an attempt to provide evidence that spreaders can largely benefit audience of news papers in social media systems, Table A.11 contrasts the number of followers of the news media profiles and the sum of the number of followers of the spreaders of each news source. Although these results do not quantify exactly the extent to which spreaders are able to increase the audience size of news sources, it clearly shows that they play a very important role in many news source audiences. For example, the number of followers in our sample of spreaders from The New York Times contains more than double the number of followers of The New York Times.

We move onto demographic of the followers of news spreaders. First, we collected followers from a sample of 25% of spreaders from our dataset. For this data sample, the average confidence level for the number of the followers of the spreaders is 6111.154 ± 66396.94, with a confidence interval of 95%. After that, we analyze the demographic characteristics of the followers of the spreaders.

Table A.12 shows the demographics of the followers of news spreaders. Compared with the demographics of the followers of news sources (Table A.5), we observe the increase in the percentage of Female–the average increase is 9%. Besides that, for race, the percentage of White is higher–the average increase is 16%. We tried to test whether this difference in demographics of spreaders' followers and those of the original followers is statistically significant.

We define the demographic distribution of the audience for each news media as a six-long vector whose element is a proportion of each demographic group (e.g., Male-Asian, Female-Asian, ..., and Female-White), respectively. With these vectors, we use the Kolmogorov-Smirnov test [162], which is a widely used statistical test to check whether two distributions are generated from an identical reference distribution. However, the difference is not statistically significant (for The New York Times, D = 0.5, p-value = 0.1641). The main reason is that the length of the vector, six, is too

Table A.12: Demographic characteristics of each the followers of the spreaders by news source.

News Media	Race (%)	Gender (%)		Total:
		Male	Female	
The New York Times	Asian	4.8	5.6	10.4
	AF-AM	6.3	4.2	10.5
	White	41.5	37.5	79.1
	Total:	52.7	47.3	100.0
Reuters	Asian	4.8	5.4	10.2
	AF-AM	6.3	4.0	10.4
	White	42.3	37.1	79.4
	Total:	53.4	46.6	100.0
The Guardian	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.8	9.9
	White	42.7	37.2	80.0
	Total:	53.6	46.4	100.0
The Wall Street Journal	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.9	10.0
	White	43.0	36.9	79.9
	Total:	54.0	46.0	100.0
BBC News	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.8	9.8
	White	42.9	37.2	80.1
	Total:	53.7	46.3	100.0

short to get statistical evidence. In future work, we will build demographic vectors for multiple snapshots and compute the statistical significance by concatenating those vectors.

A.6.2 News Popularity and Demographics

In the previous section, we show that understanding news spreaders is important as they extend the readership of news media. Another important aspect is whether the demographic traits of news spreaders are relating to the popularity of news. To this end, we collect the number clicks for each news URL using the Bit.ly API[13]. Then, we compare the popularity of news articles shared by different demographic groups to know whether a certain demographic group share news URLs likely to be more popular.

For gender group, we observe that the news items shared by Female are more clicked that those shared by Male. The differences are statistically significant by Kruskal-Wallis H-test ($H = 7.719, p < 0.005$). For race, the news articles shared

[13] https://dev.bitly.com/

by Asians are more clicked ($H = 6.659, p < 0.005$). The results show that the demographic information of news spreaders can potentially help in predicting the popularity of news articles.

A.7 Concluding Discussion

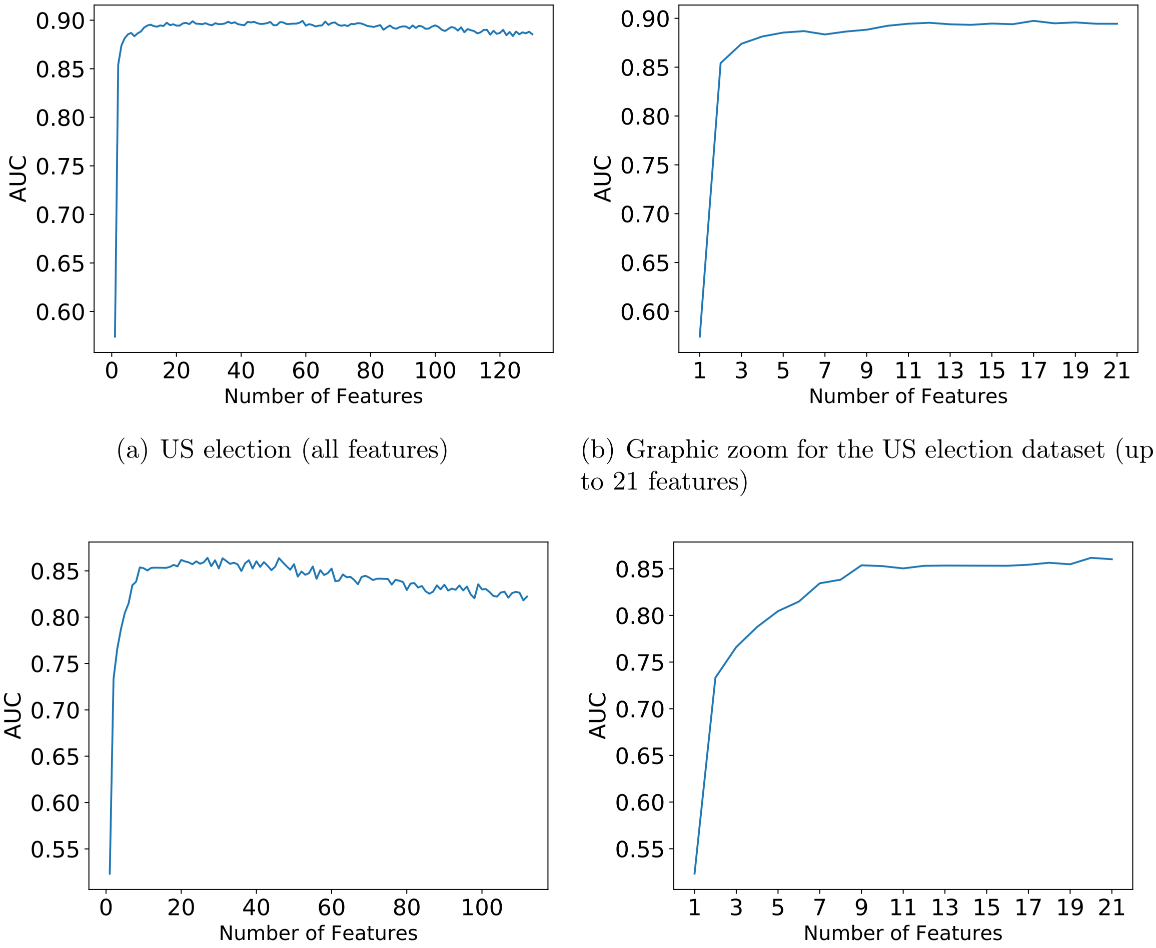
The increasing diffusion of news in social media systems, associated with the great power provided to users along the dissemination process, are making these platforms a fertile ground for misleading or fake news propagation. The growing use of Twitter as a news' channel highlights the importance of characterizing news spreaders to understand who they are, what they share and their impact. Next, we briefly discuss implications of our main findings and discuss directions we aim to explore next.

Bias on breaking news stories: A widely used tool that users use to find breaking news-stories in online social networks is the Trending stories (or topics) [80, 269]. Recently, Facebook has been involved in many controversies related to trending stories [70]. First, Facebook involved human curators as part of its process to identify trending stories. A main criticism was that human curators could bias the final list of stories. Then, Facebook removed the human intervention and followed the popular perception that data-driven algorithms would not be biased as they simply process data. Our results, however, shows the data itself is biased, at least in terms of the demographic groups considered. We show that demographic groups of white and male users tend to share more news in Twitter. Our results also quantify the existing bias on Twitter shares towards specific demographic groups across news categories and other dimensions. Thus, our work contributes with a new and important perspective to the emerging debate in the community centered around concerns about bias and transparency of decisions taken by algorithms operating over user-generated data. Finally, we believe that the increasing availability of information about demographics will help the development of systems that promote more diversity and less inequality to users. Thus, as a final contribution of our effort, we intend to release our demographic dataset to the research community by the time of publication of this study.

Personalized news recommendations: Our analysis shows different user behaviors in terms of news sharing and also highlight demographic differents in terms of user interests. Identifying intrinsic characteristics of the users who spread the news in the online world and identifying how users interest across demographics is a key step towards the development of a framework that can promote the customization of the user experience using social media for news digest. We aim at further exploring this topic as part of our future work by investigating the discriminative power of demographic, linguistic, and network features in predicting a user's interest in specific news and news topics.

Appendix B

Accuracy and Number of Features in the Models



(a) US election (all features)

(b) Graphic zoom for the US election dataset (up to 21 features)

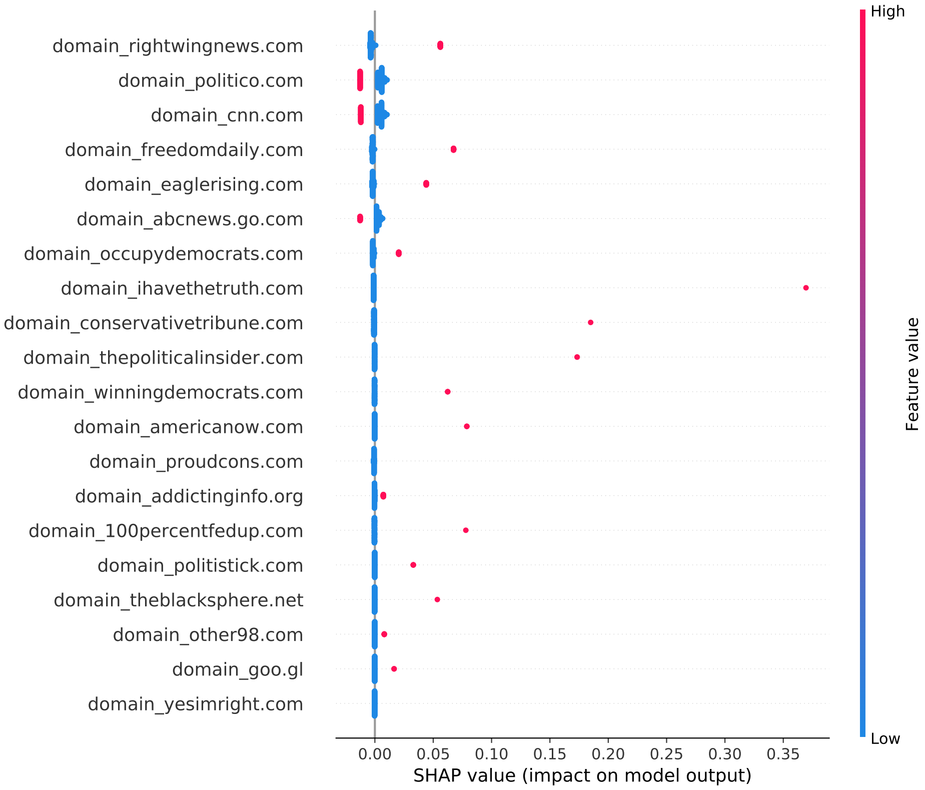
(c) Brazilian election (all features)

(d) Graphic zoom for the Brazilian election dataset (up to 21 features)

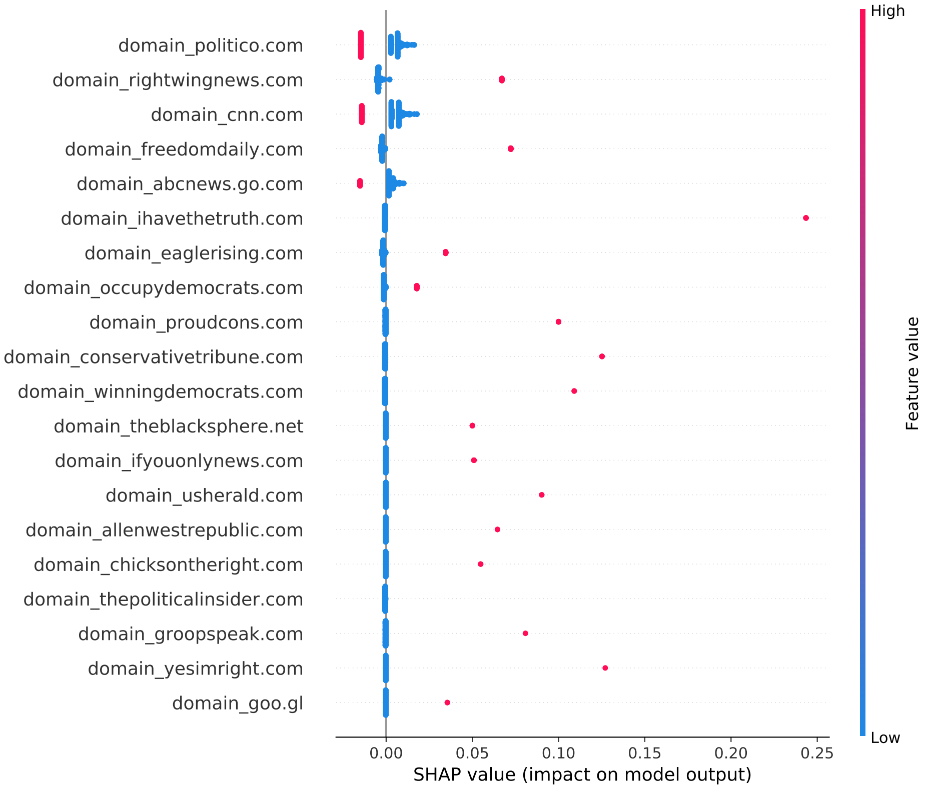
Figure B.1: Performance of XGB models in terms of AUC considering the number of features used by them. For both datasets, the predictive accuracy improves for models composed of up to 20 features. Thereafter we note a stabilization or worsening of the performance of models.

Appendix C

SHAP Graphs

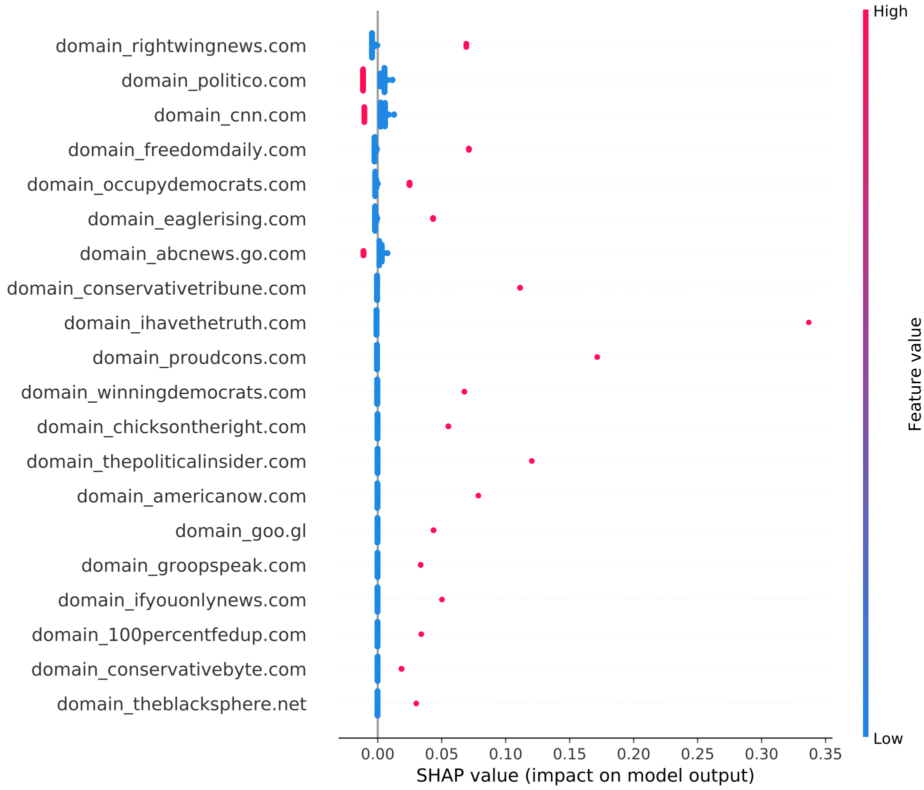


(a) Fold 1

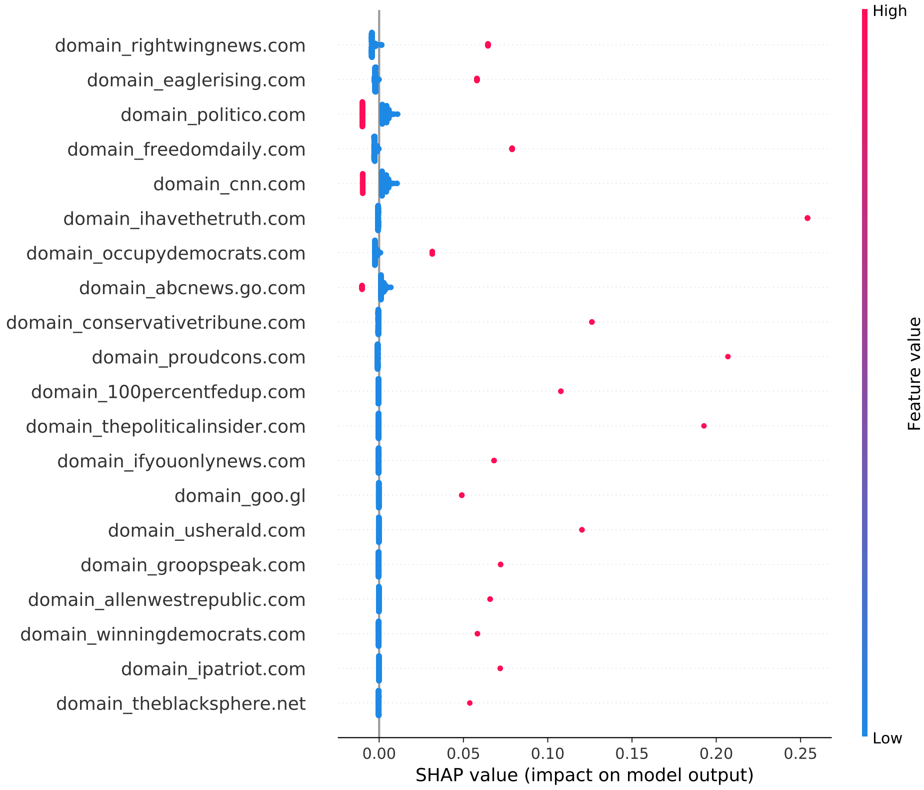


(b) Fold 2

Figure C.1: SHAP summaries for the closest models to **cluster 1** centroid for the US election dataset (Folds 1 and 2).

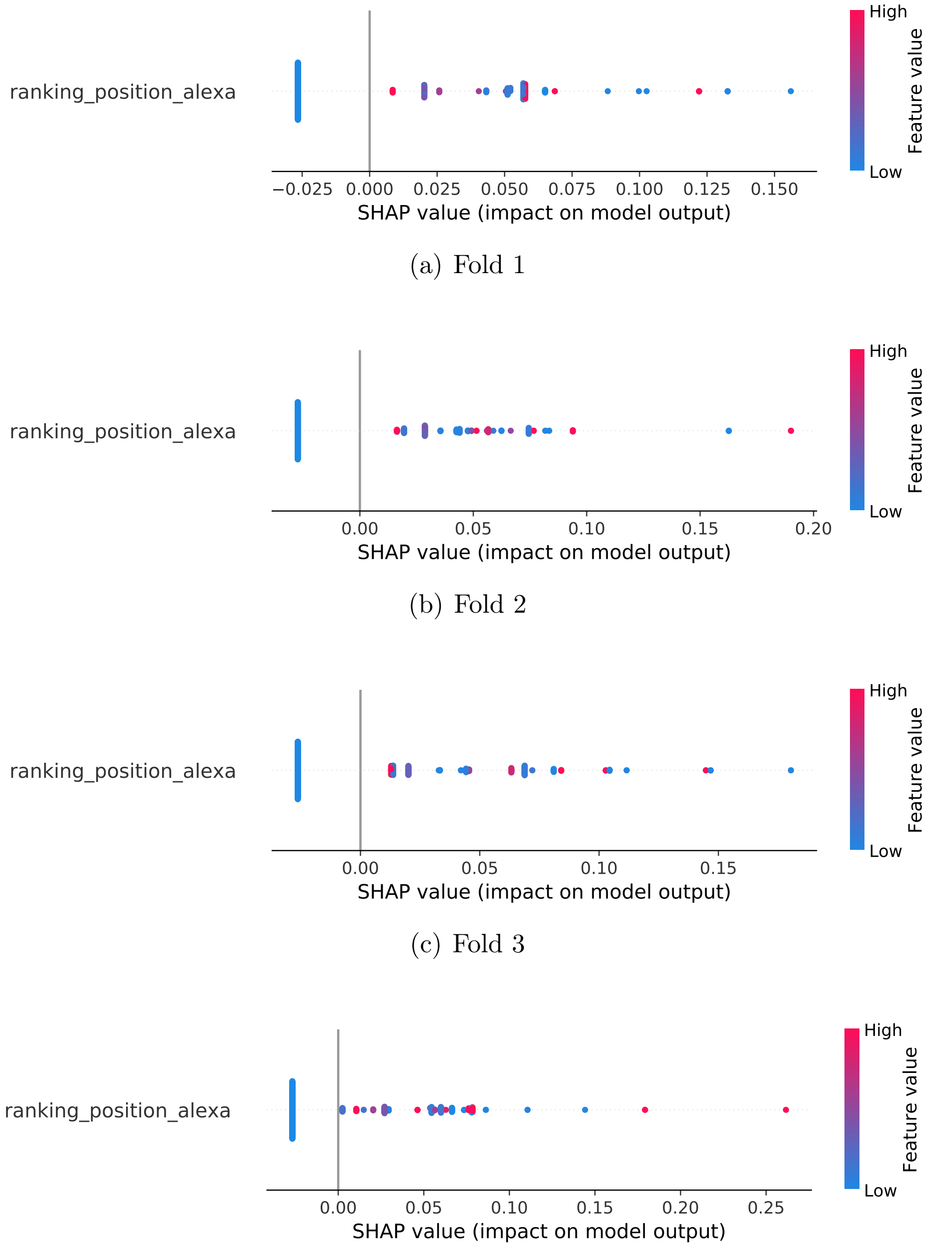


(a) Fold 3



(b) Fold 4

Figure C.2: SHAP summaries for the closest models to **cluster 1** centroid for the US election dataset (Folds 3 and 4).



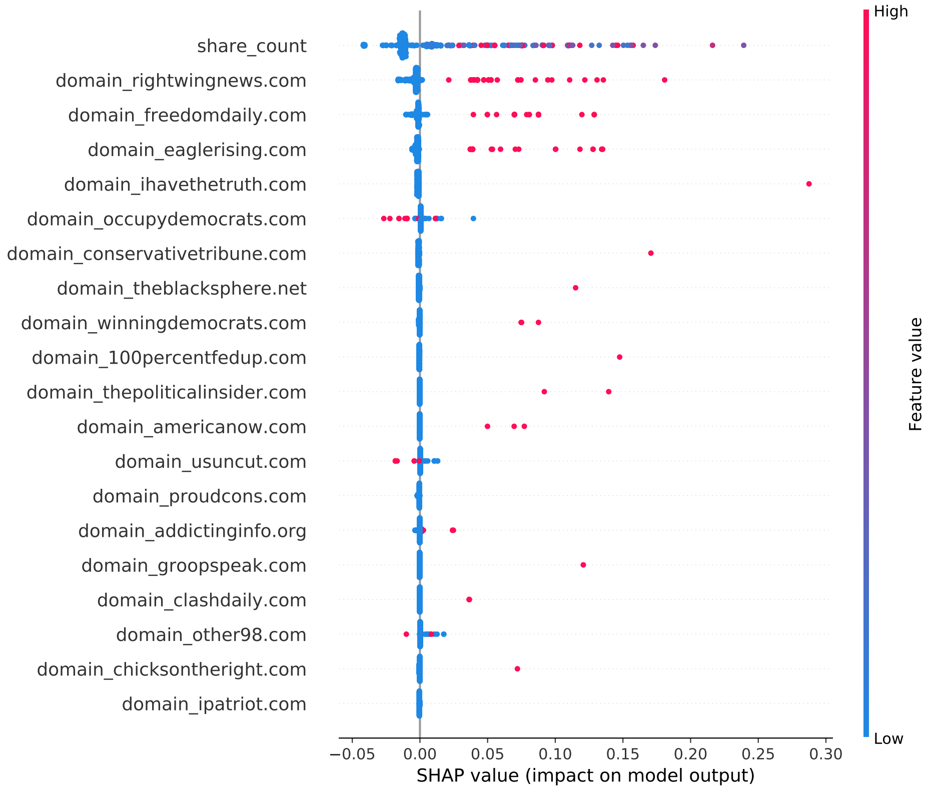
(a) Fold 1

(b) Fold 2

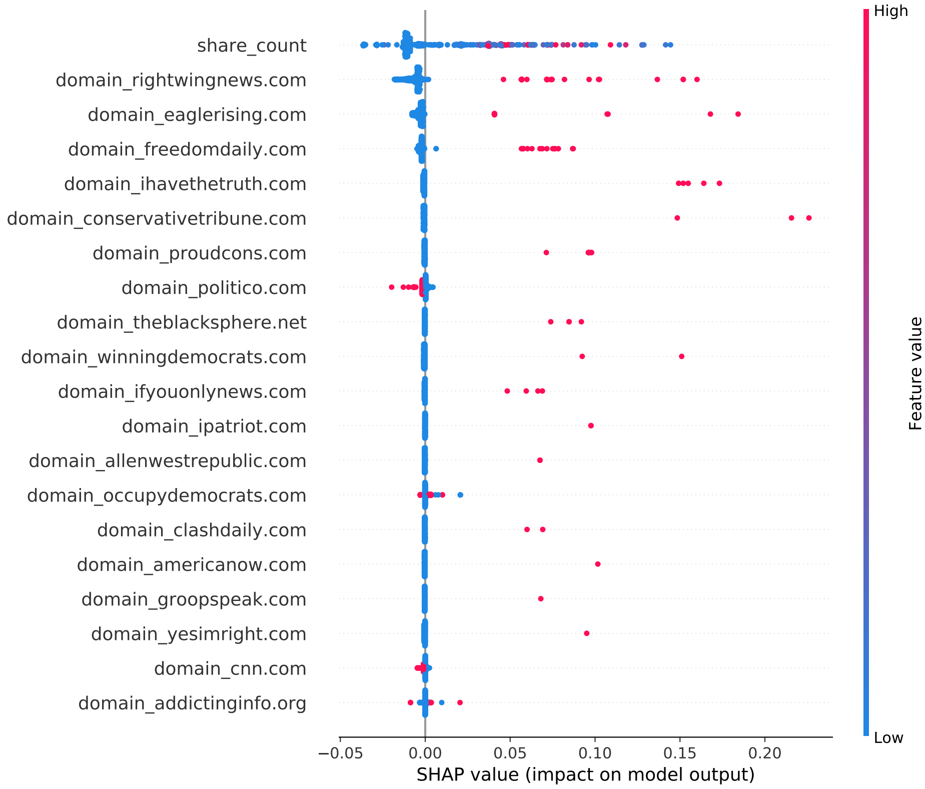
(c) Fold 3

(d) Fold 4

Figure C.3: SHAP summaries for the closest models to **cluster 2** centroid for the US election dataset (Folds 1, 2, 3 and 4).

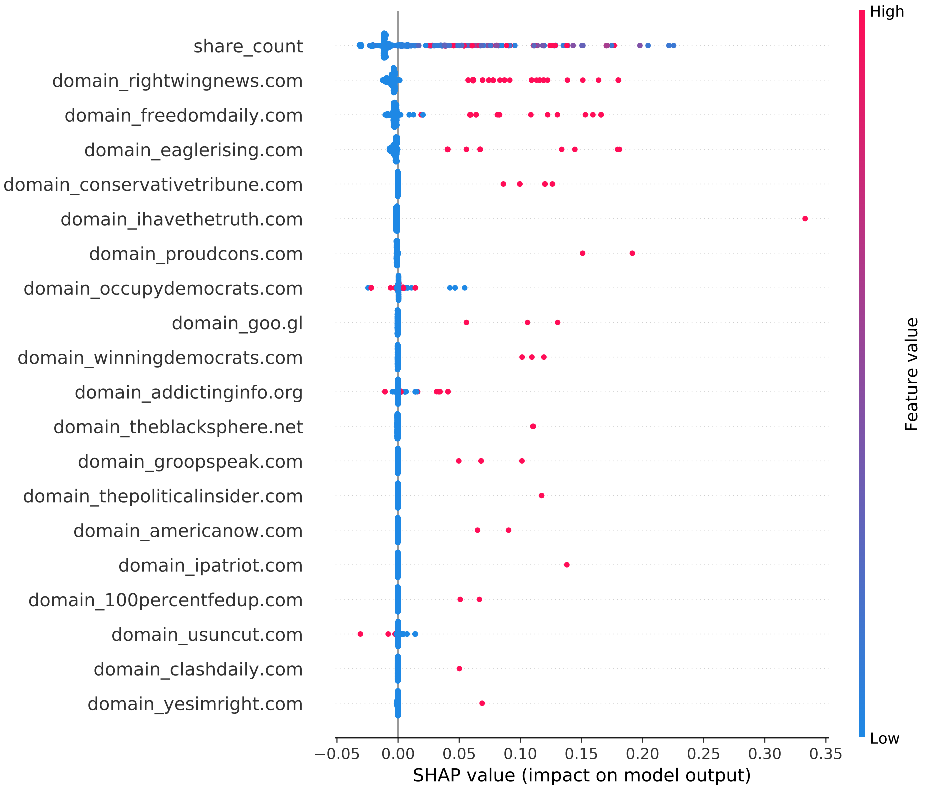


(a) Fold 1

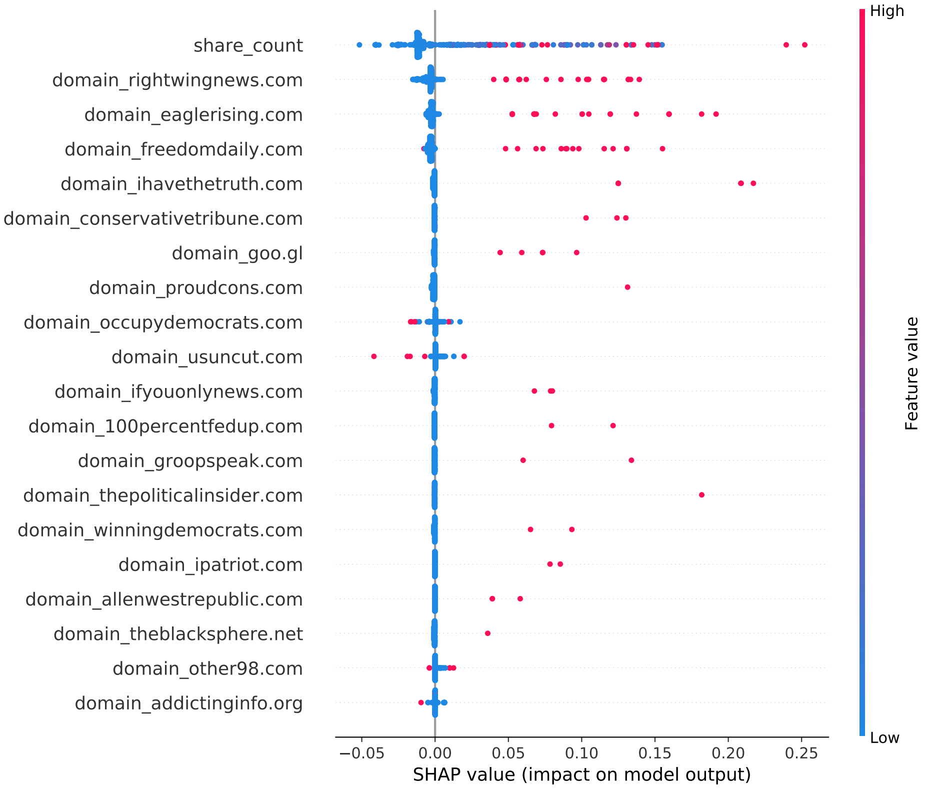


(b) Fold 2

Figure C.4: SHAP summaries for the closest models to **cluster 3** centroid for the US election dataset (Folds 1 and 2).

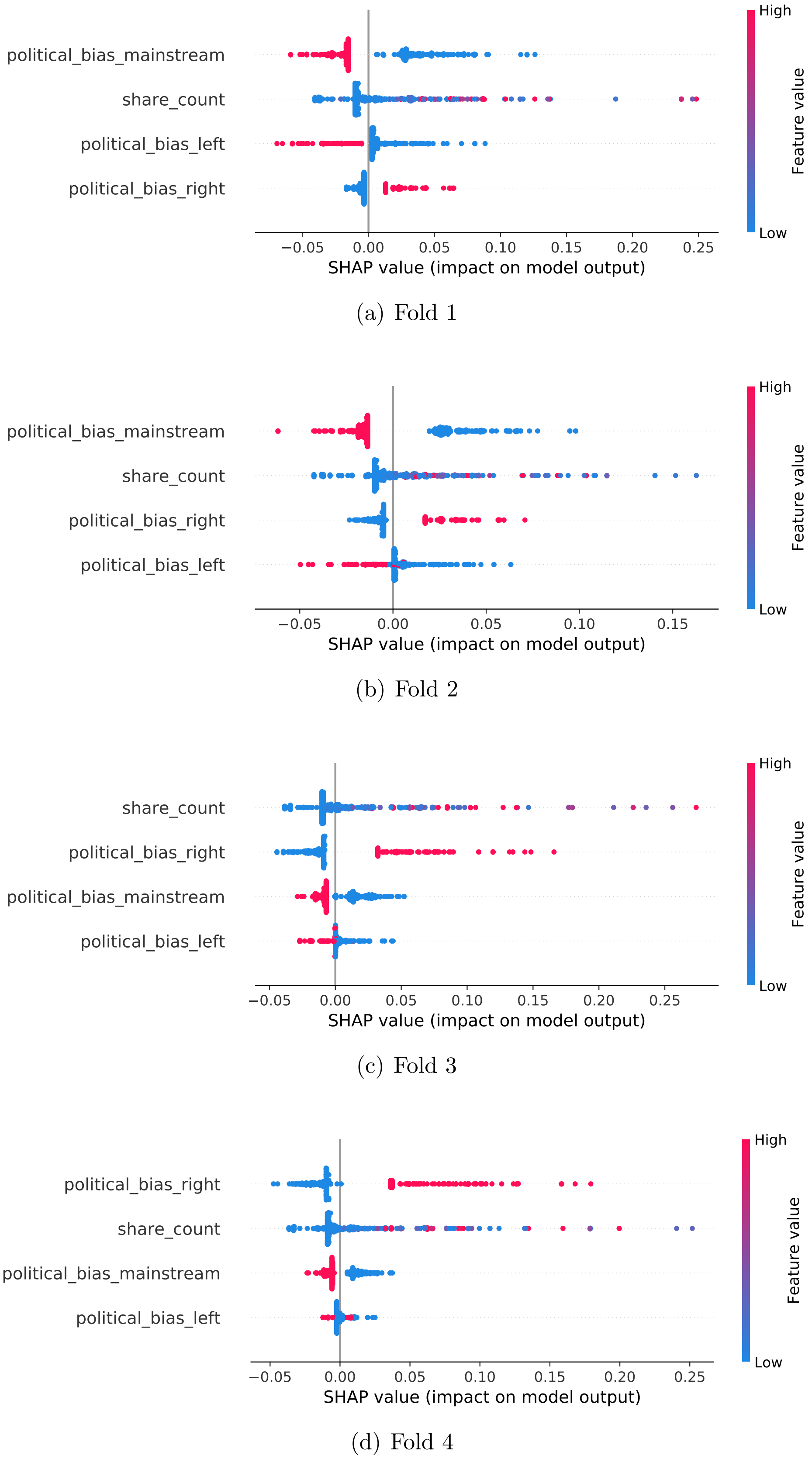


(a) Fold 3



(b) Fold 4

Figure C.5: SHAP summaries for the closest models to **cluster 3** centroid for the US election dataset (Folds 3 and 4).



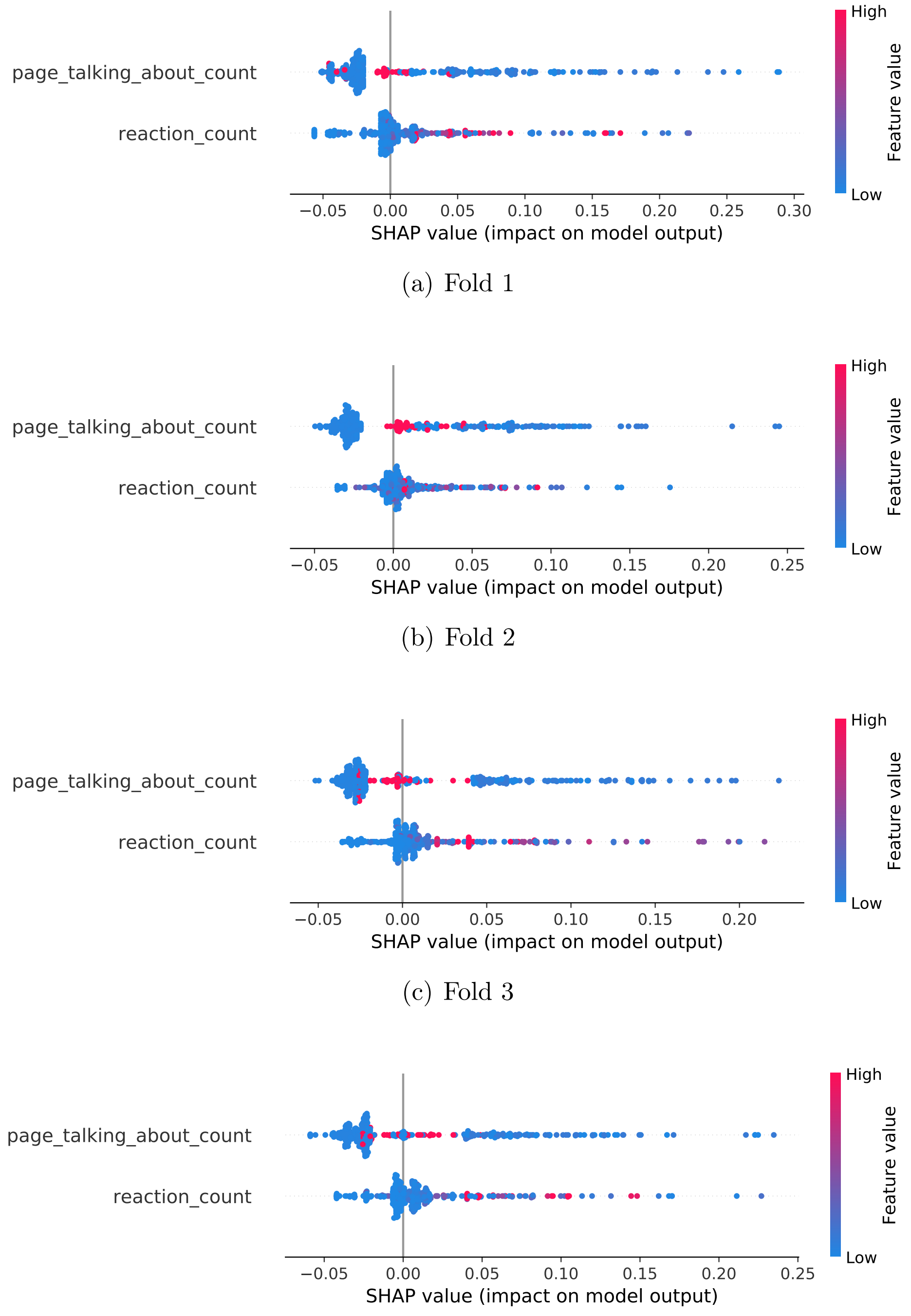
(a) Fold 1

(b) Fold 2

(c) Fold 3

(d) Fold 4

Figure C.6: SHAP summaries for the closest models to **cluster 4** centroid for the US election dataset (Folds 1, 2, 3 and 4).



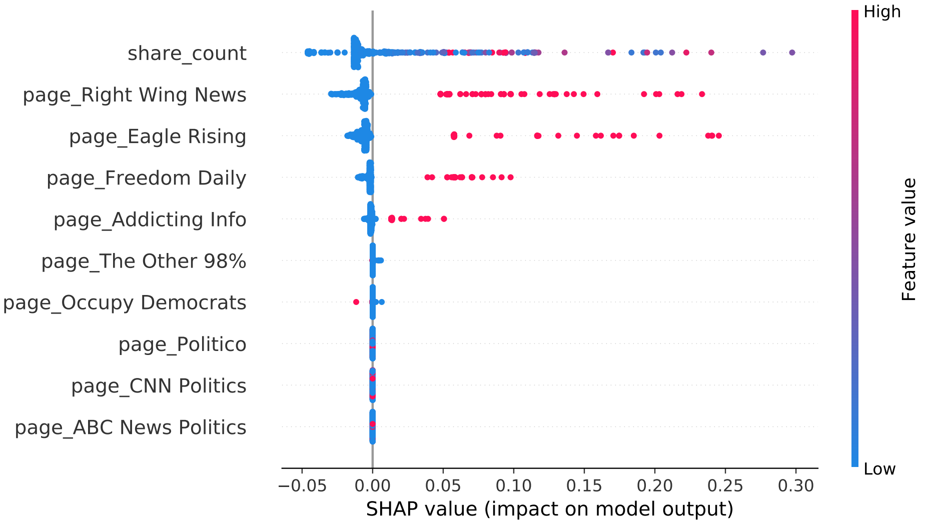
(a) Fold 1

(b) Fold 2

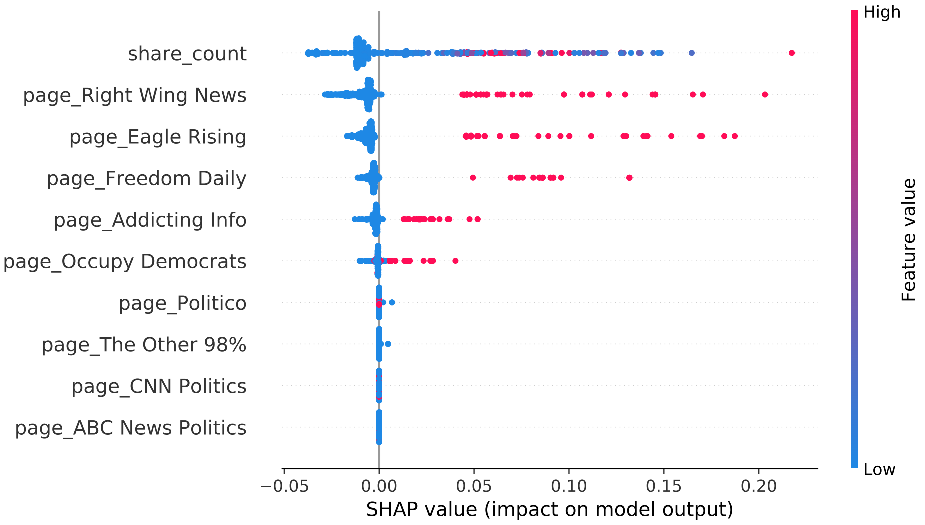
(c) Fold 3

(d) Fold 4

Figure C.7: SHAP summaries for the closest models to **cluster 5** centroid for the US election dataset (Folds 1, 2, 3 and 4).

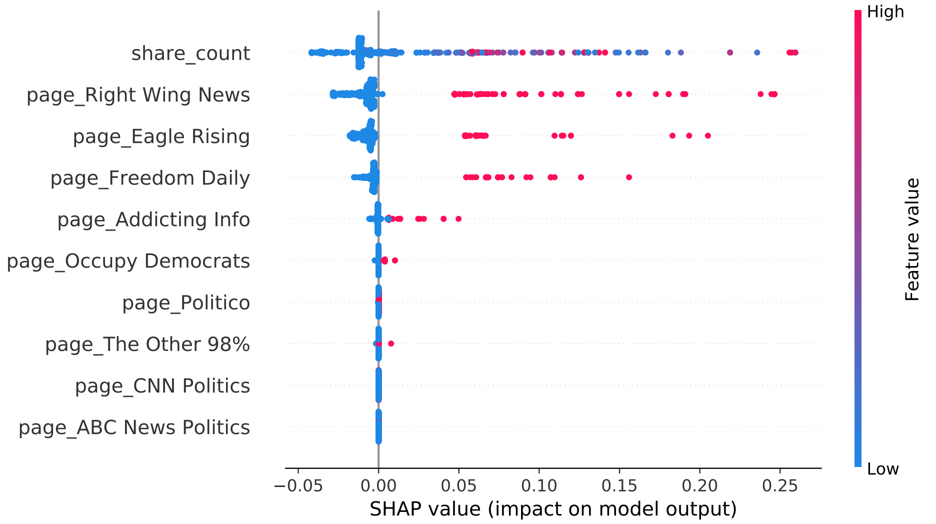


(a) Fold 1

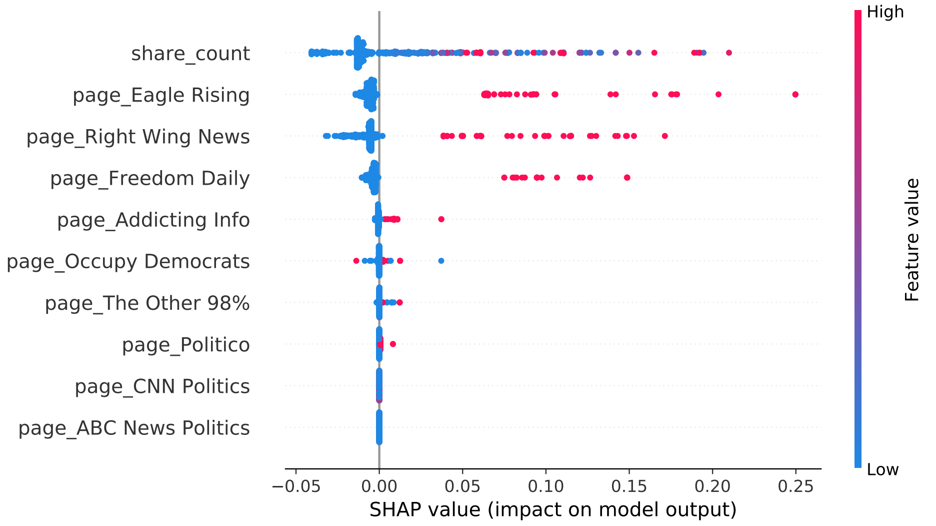


(b) Fold 2

Figure C.8: SHAP summaries for the closest models to **cluster 6** centroid for the US election dataset (Folds 1 and 2).

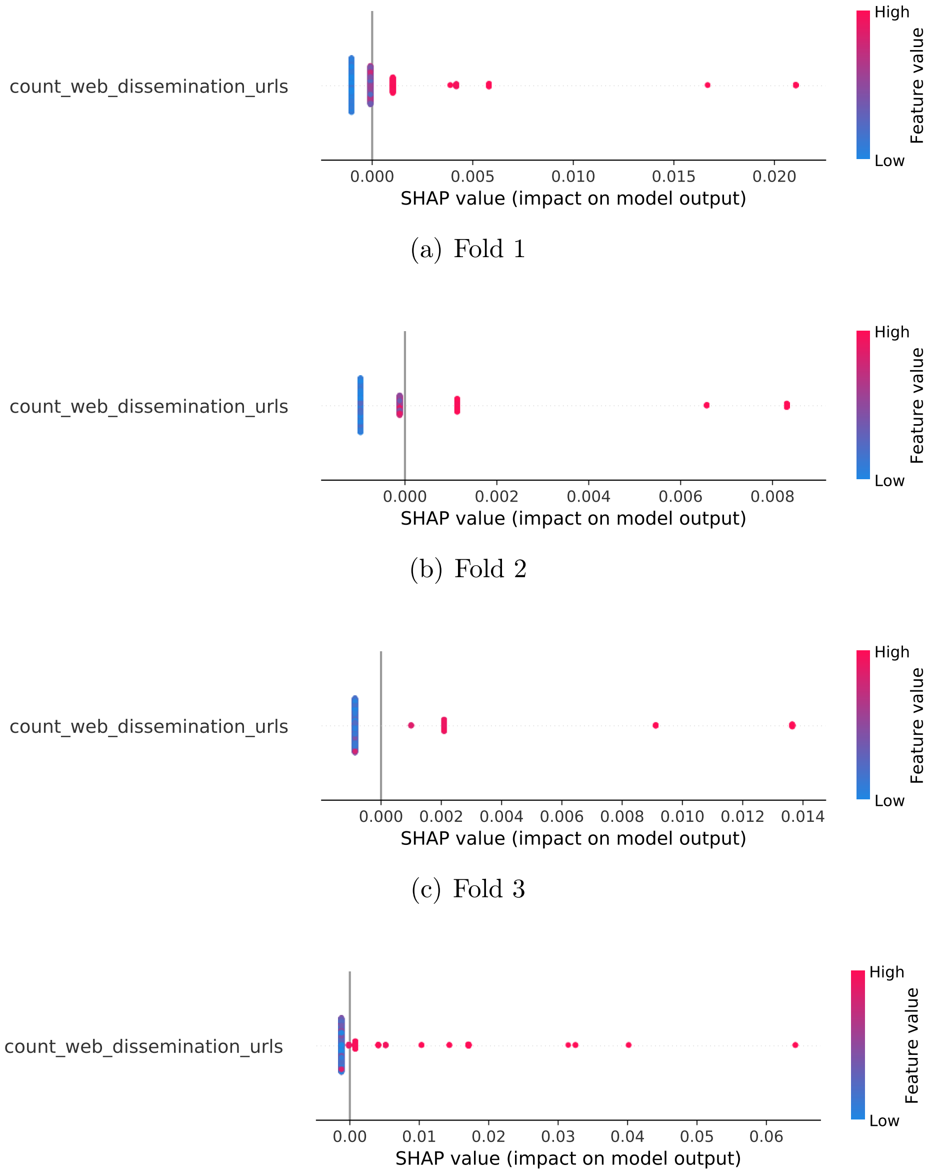


(a) Fold 3



(b) Fold 4

Figure C.9: SHAP summaries for the closest models to **cluster 6** centroid for the US election dataset (Folds 3 and 4).



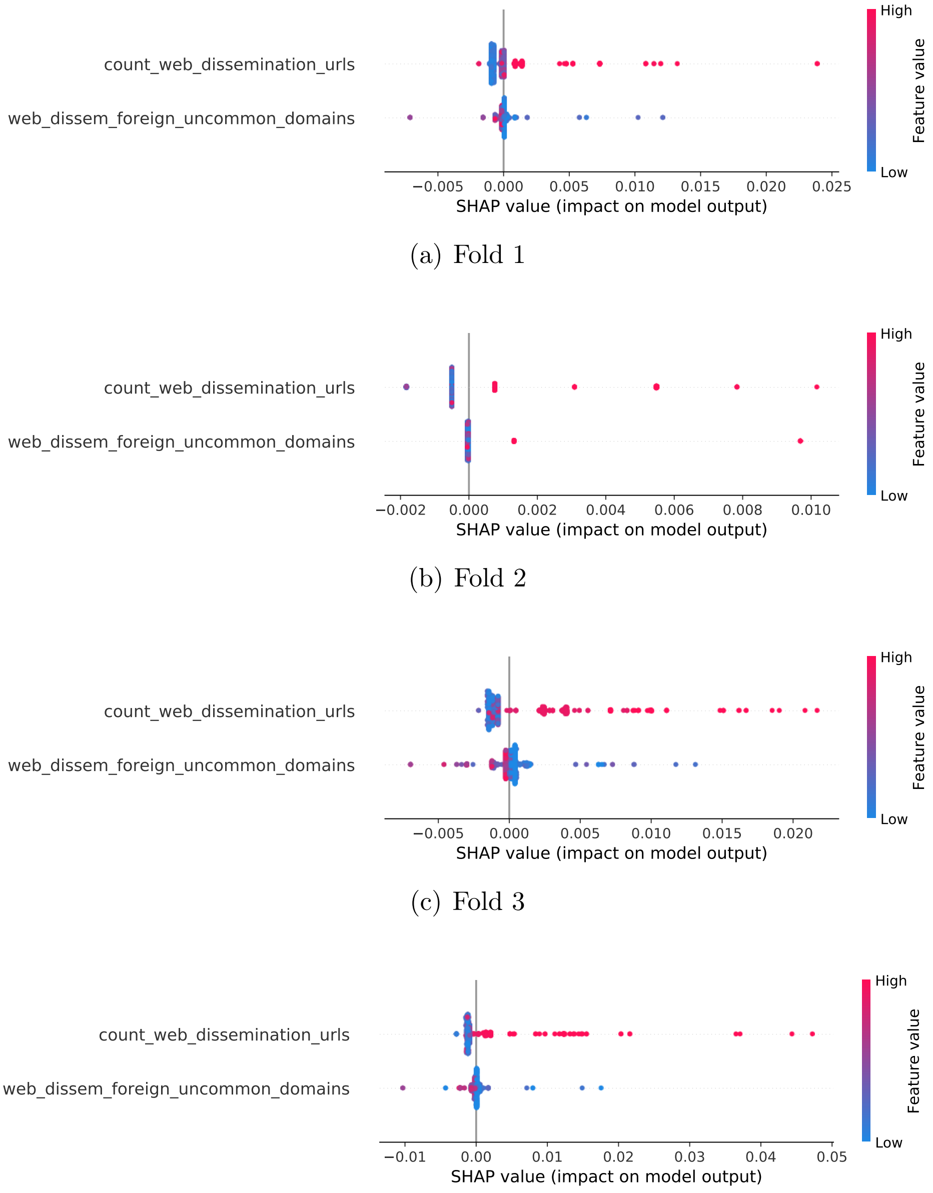
(a) Fold 1

(b) Fold 2

(c) Fold 3

(d) Fold 4

Figure C.10: SHAP summaries for the closest models to **cluster 1** centroid for the Brazilian election dataset (Folds 1, 2, 3 and 4).



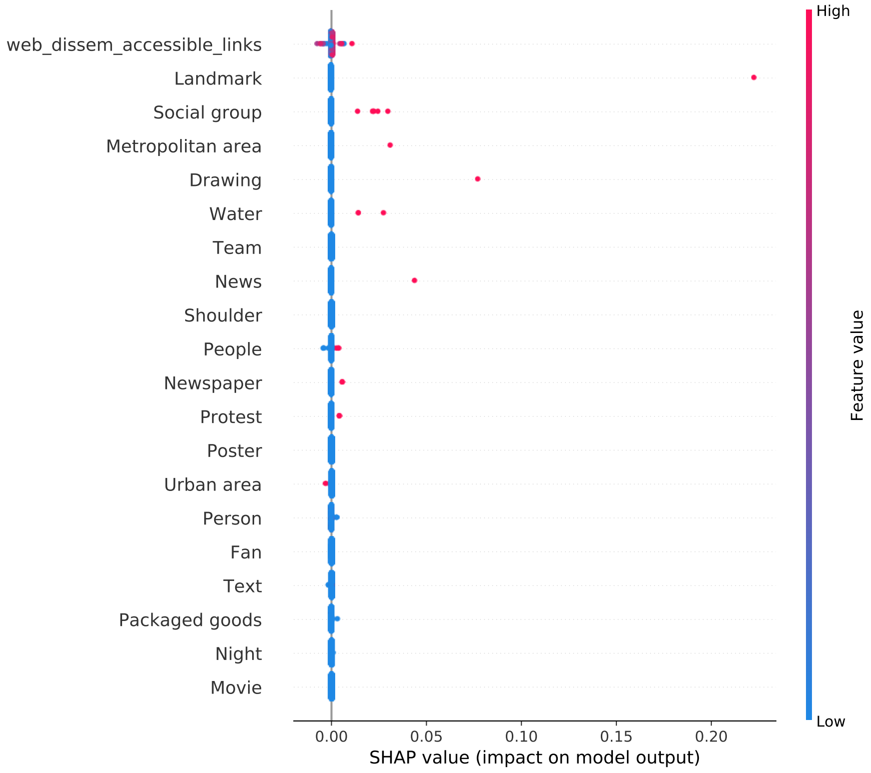
(a) Fold 1

(b) Fold 2

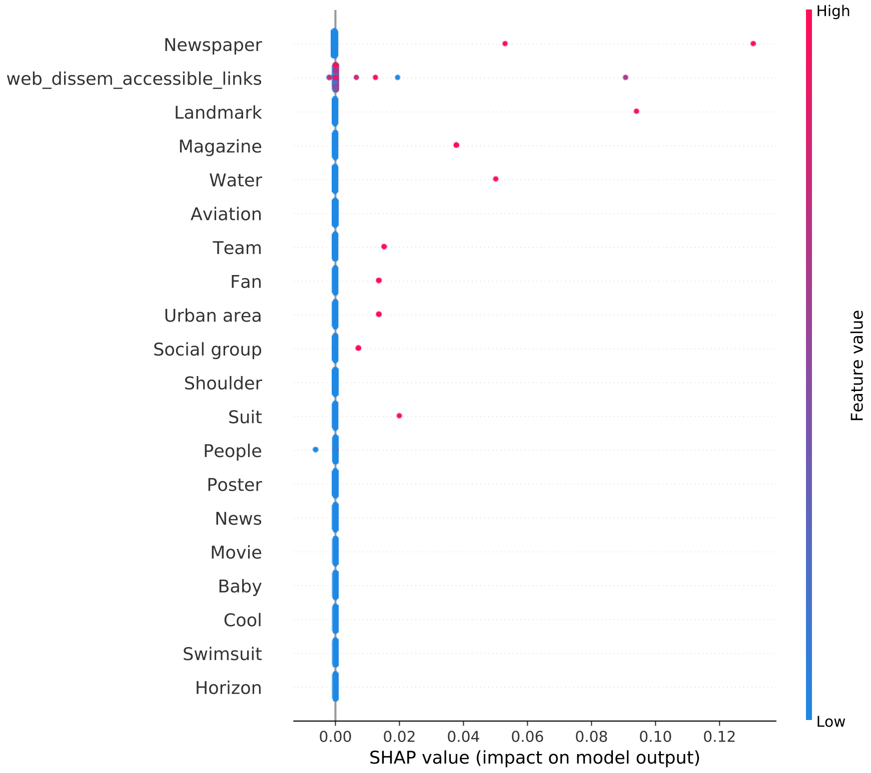
(c) Fold 3

(d) Fold 4

Figure C.11: SHAP summaries for the closest models to **cluster 2** centroid for the Brazilian election dataset (Folds 1, 2, 3 and 4).

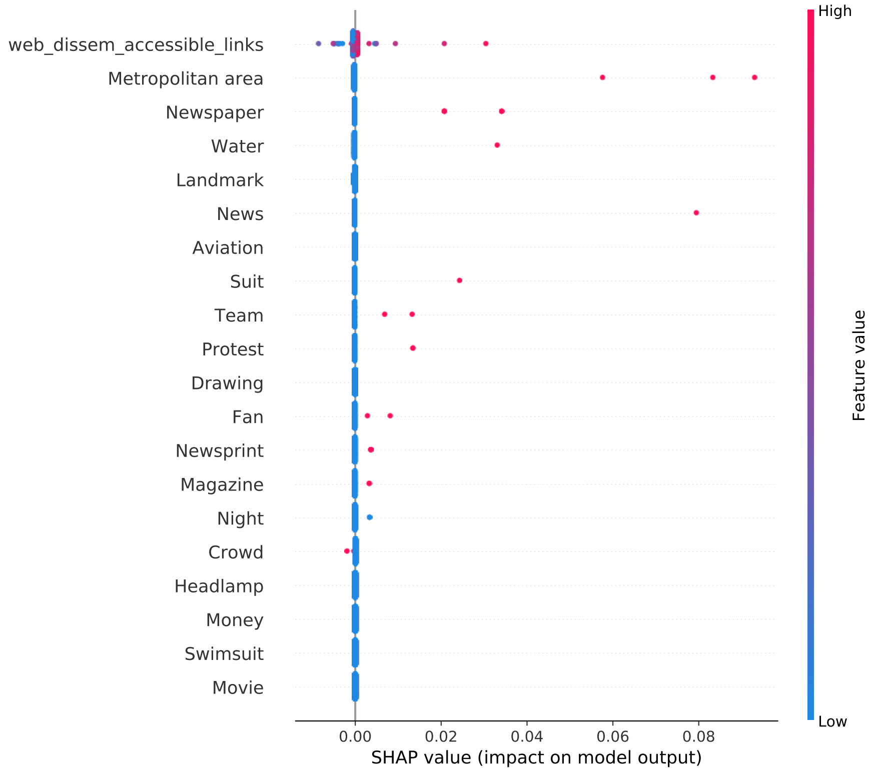


(a) Fold 1

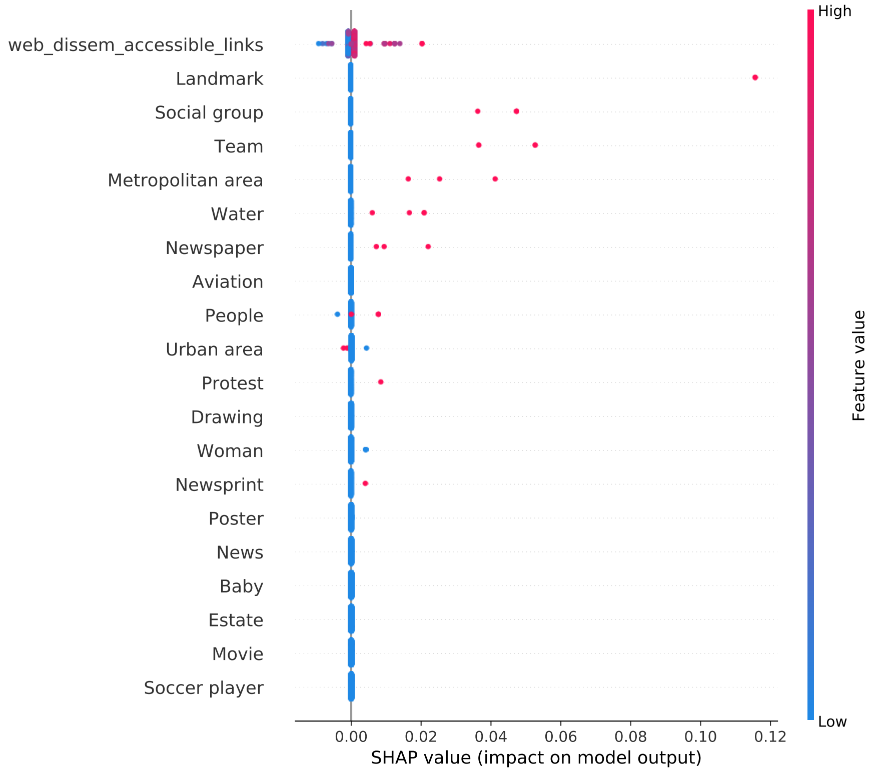


(b) Fold 2

Figure C.12: SHAP summaries for the closest models to **cluster 3** centroid for the Brazilian election dataset (Folds 1 and 2).

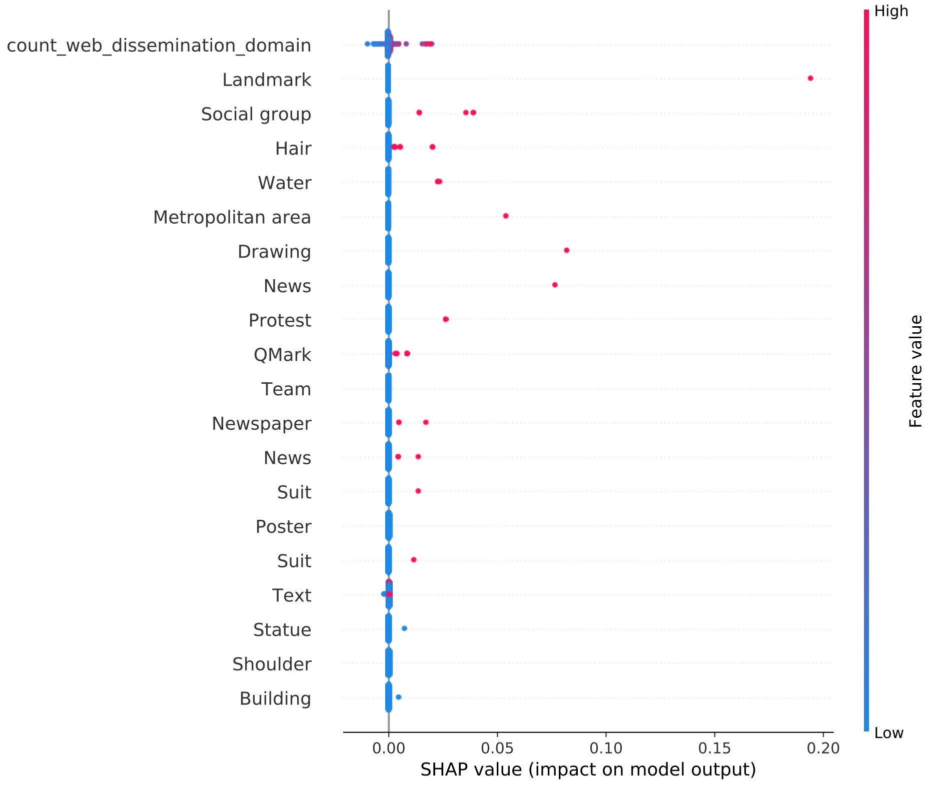


(a) Fold 3

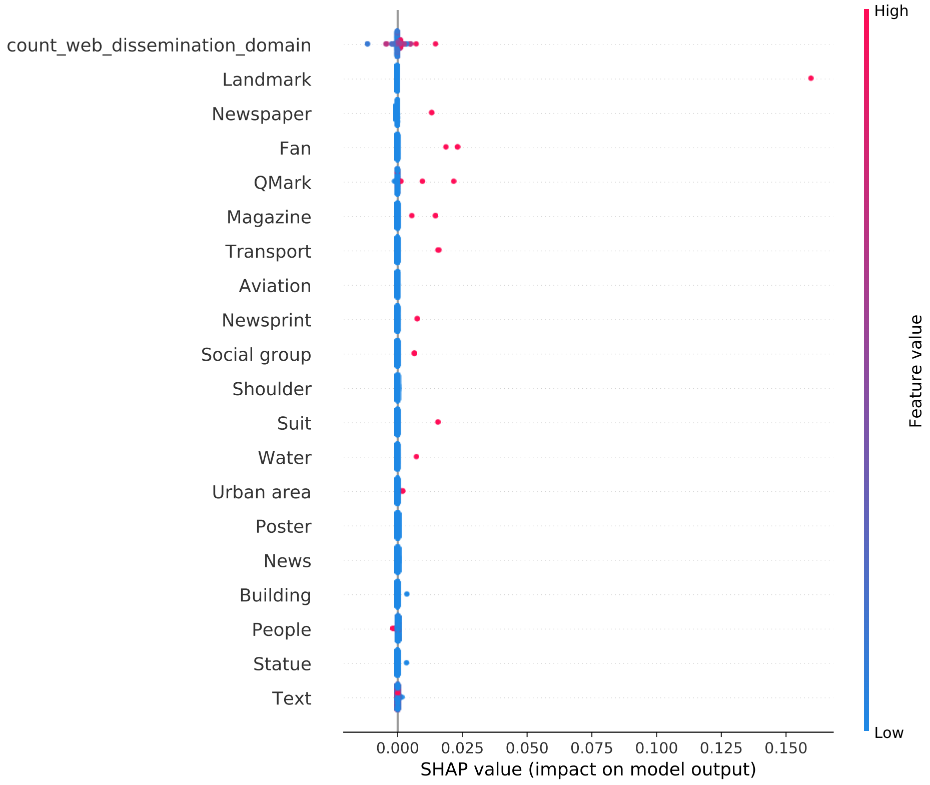


(b) Fold 4

Figure C.13: SHAP summaries for the closest models to **cluster 3** centroid for the Brazilian election dataset (Folds 3 and 4).

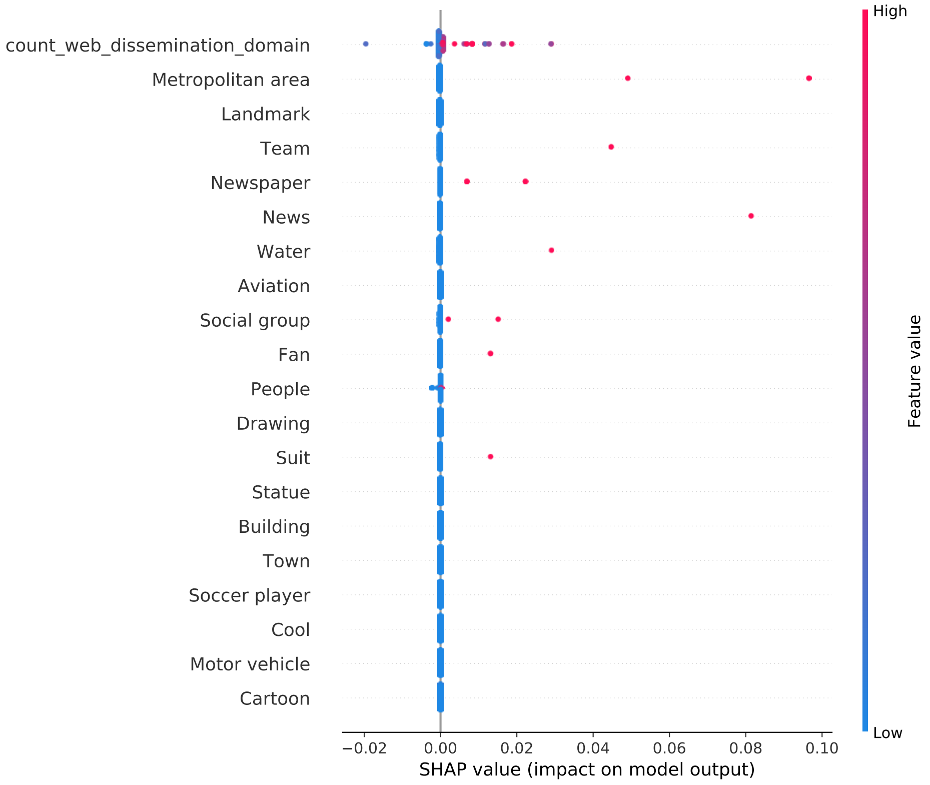


(a) Fold 1

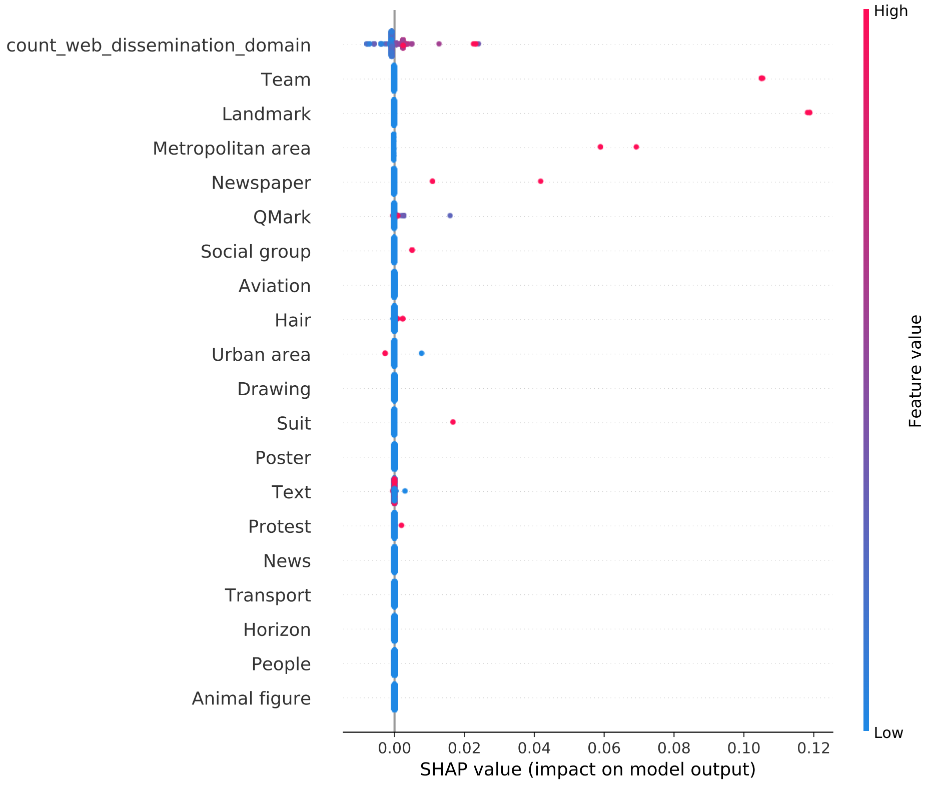


(b) Fold 2

Figure C.14: SHAP summaries for the closest models to **cluster 4** centroid for the Brazilian election dataset (Folds 1 and 2).

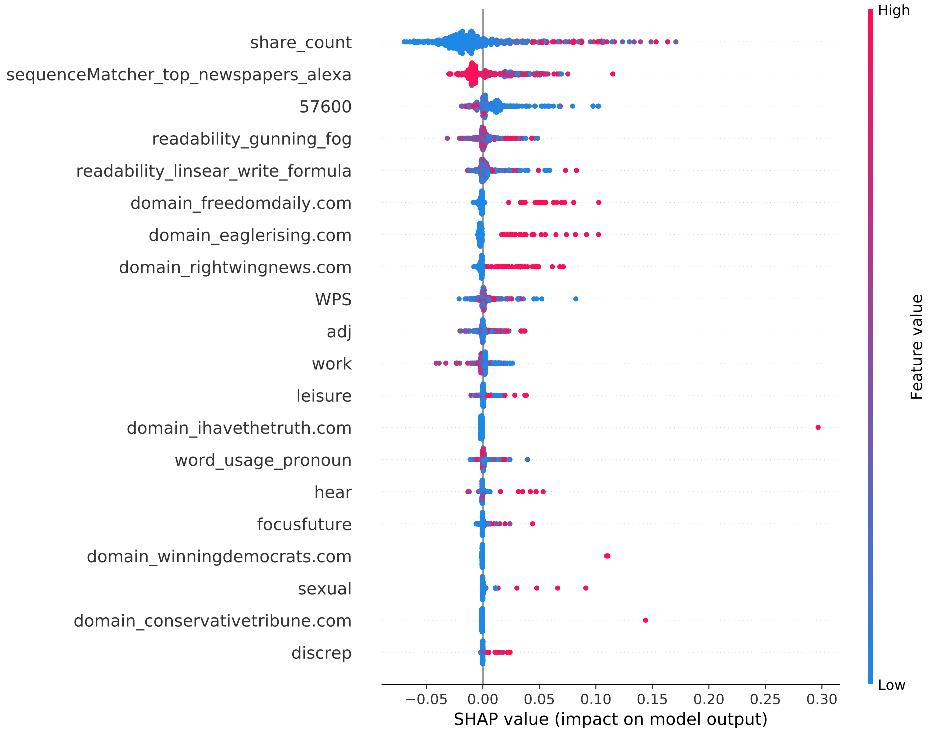


(a) Fold 3

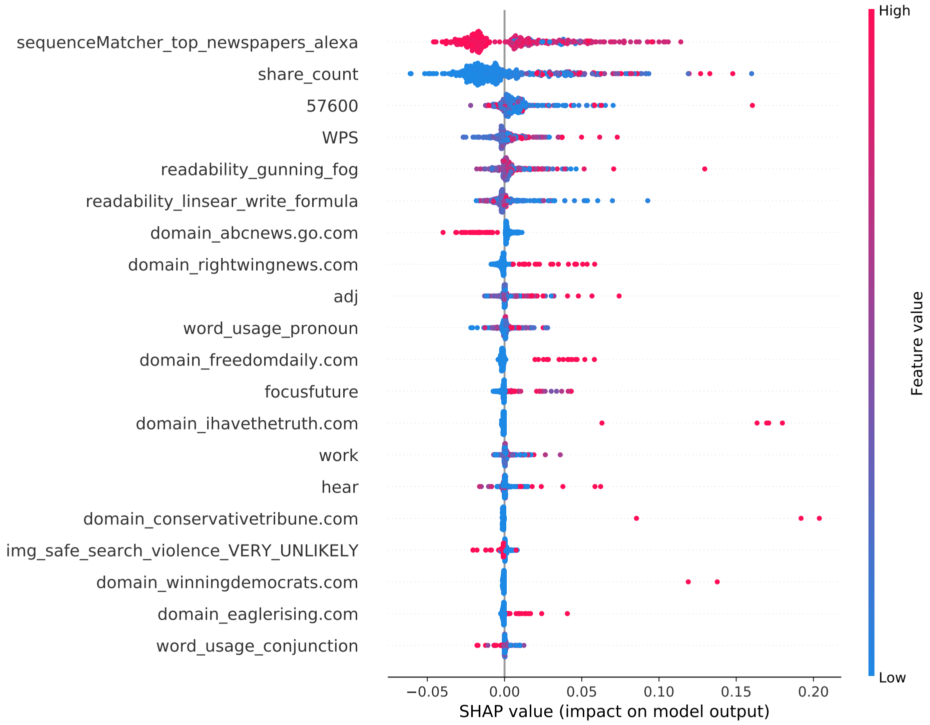


(b) Fold 4

Figure C.15: SHAP summaries for the closest models to **cluster 4** centroid for the Brazilian election dataset (Folds 3 and 4).

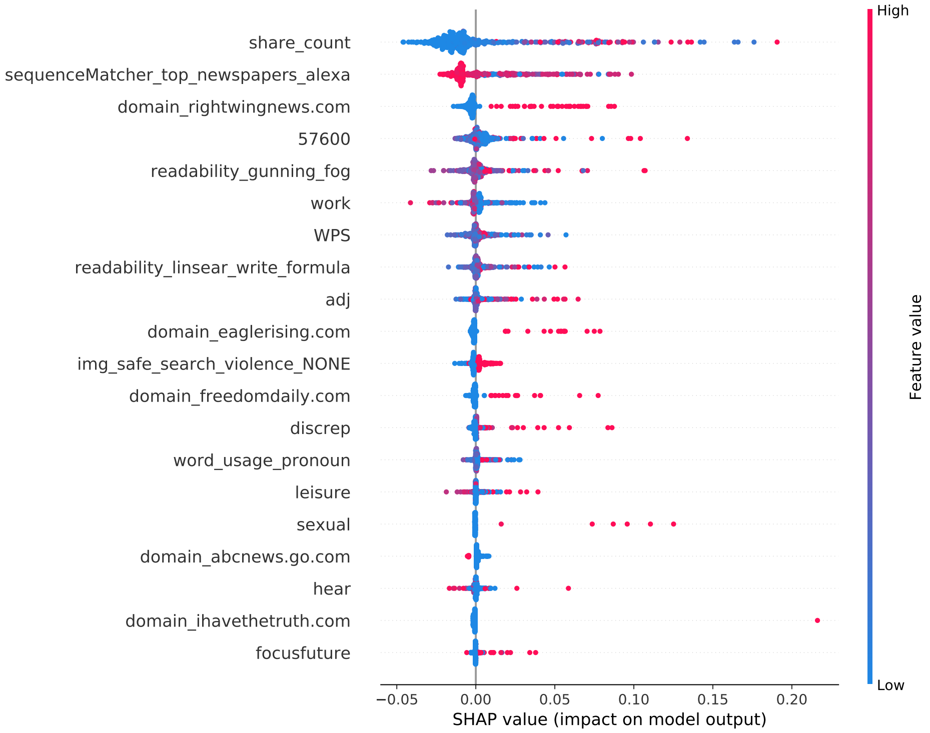


(a) Fold 1

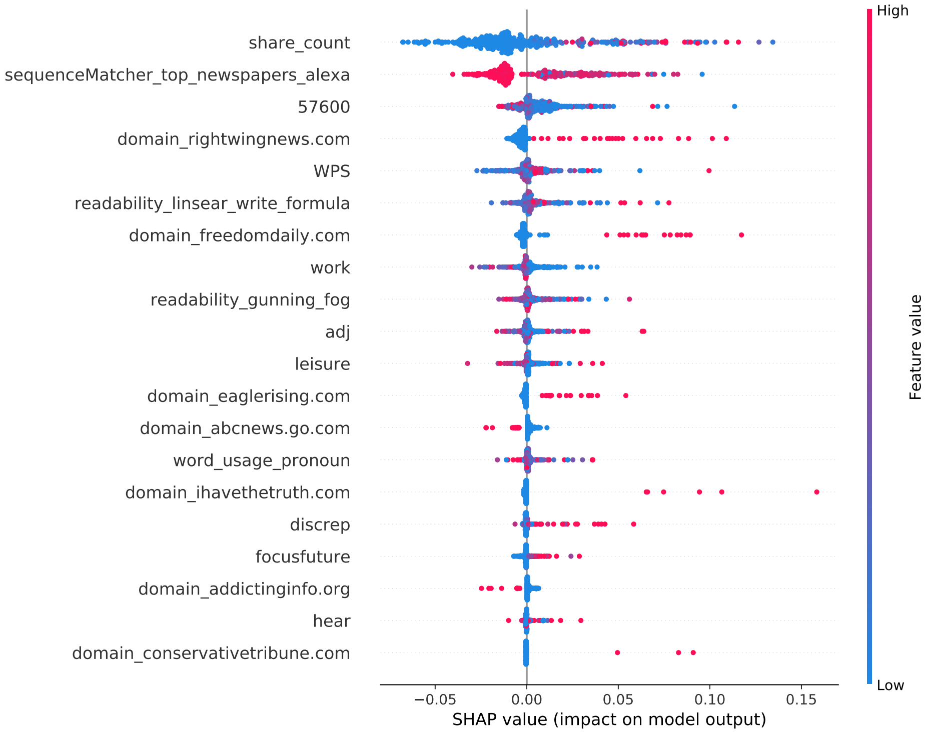


(b) Fold 2

Figure C.16: SHAP summaries for the highest AUC model in **cluster 1** for the US election dataset (Folds 1 and 2).

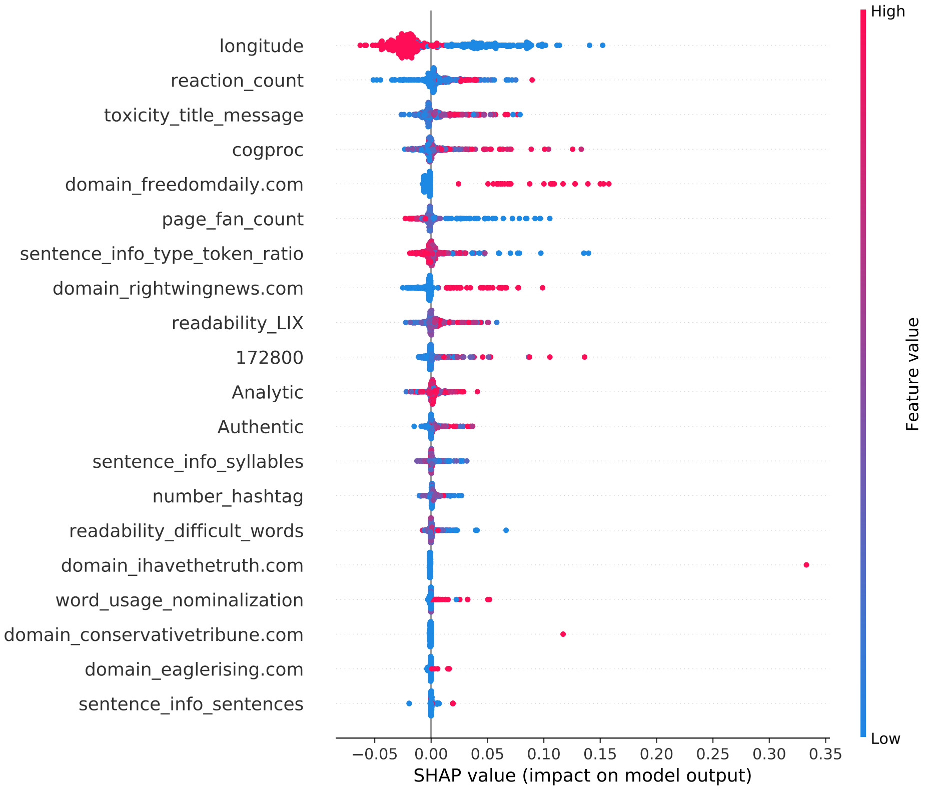


(a) Fold 3

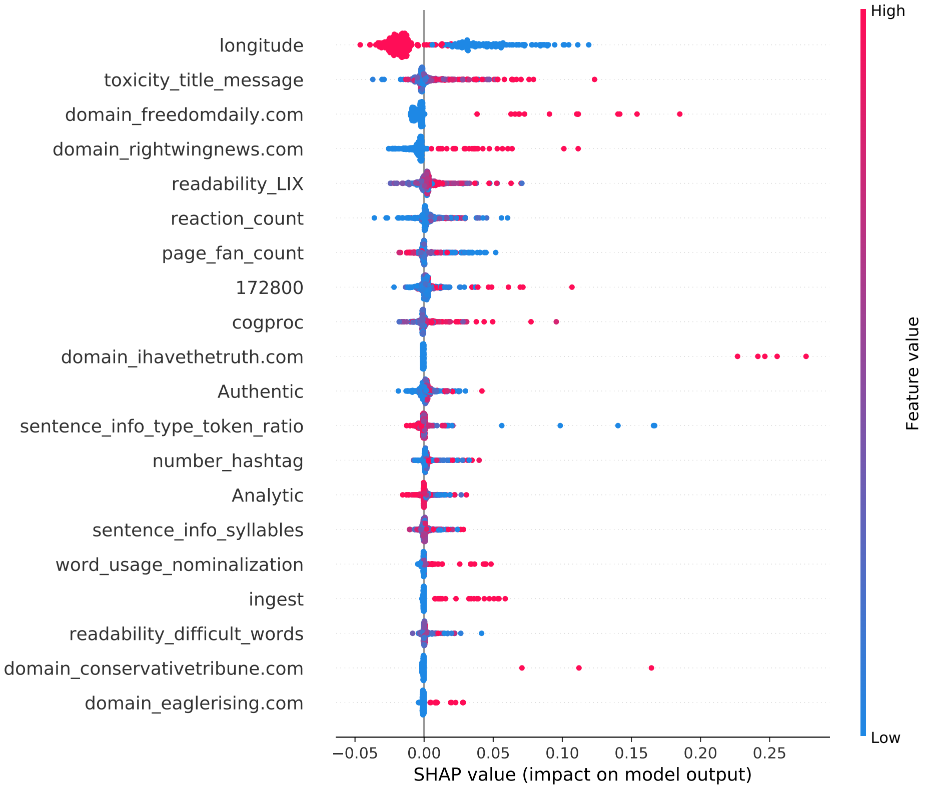


(b) Fold 4

Figure C.17: SHAP summaries for the highest AUC model in **cluster 1** for the US election dataset (Folds 3 and 4).

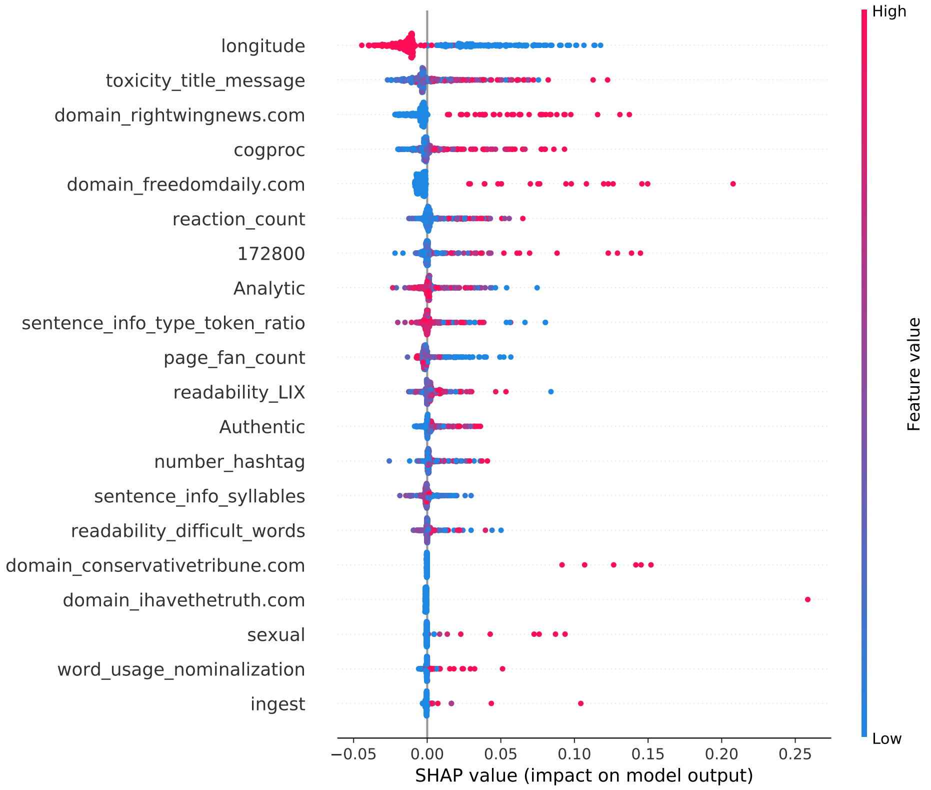


(a) Fold 1

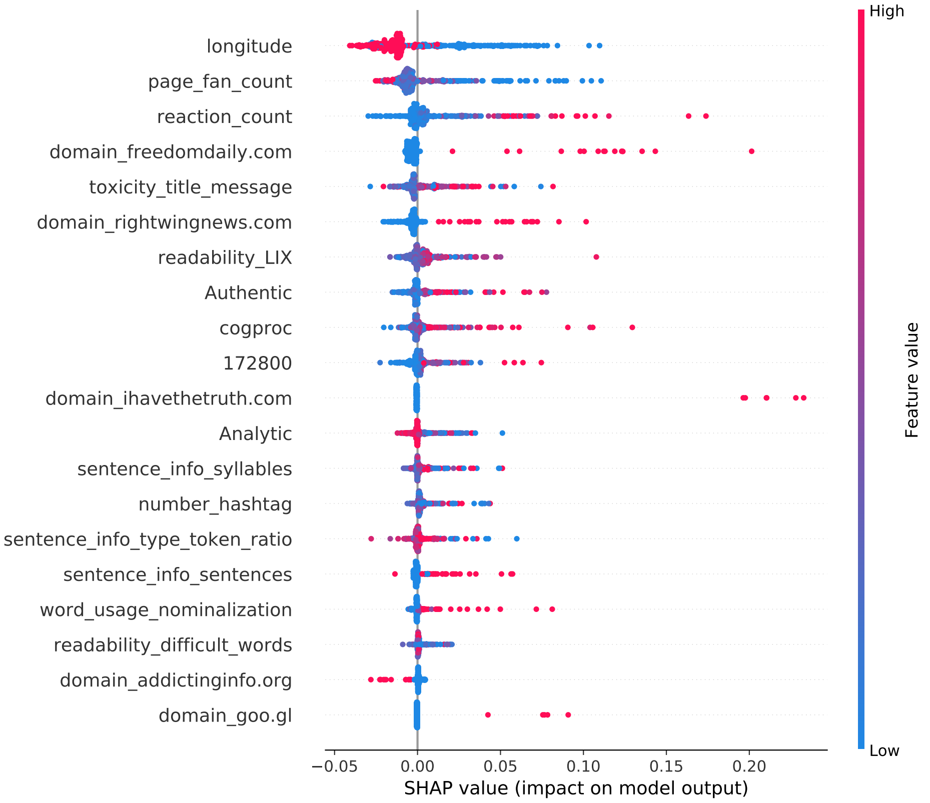


(b) Fold 2

Figure C.18: SHAP summaries for the highest AUC model in **cluster 2** for the US election dataset (Folds 1 and 2).

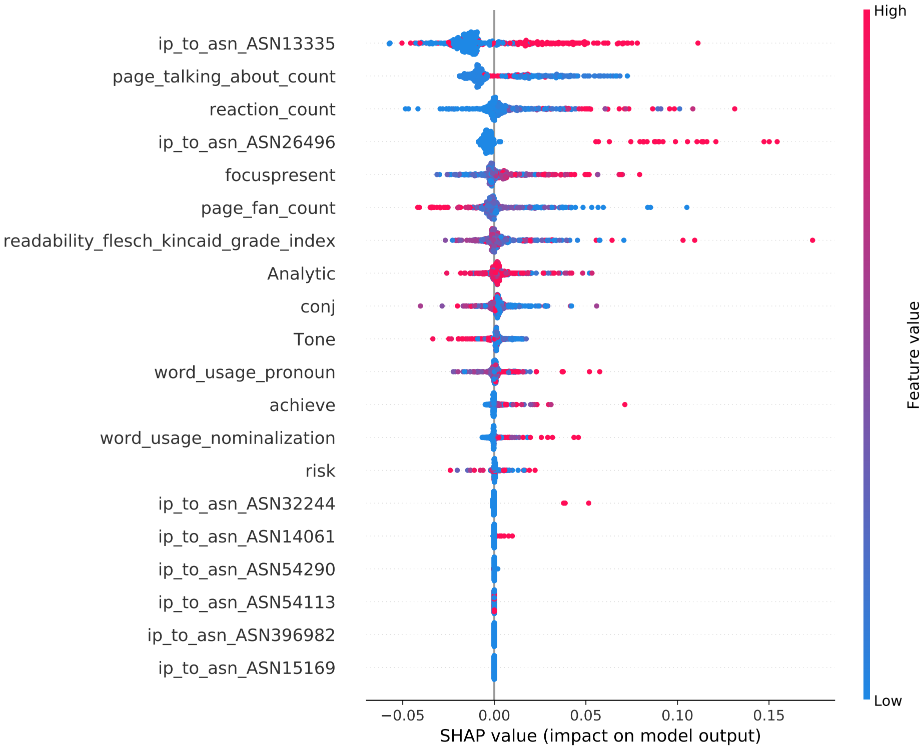


(a) Fold 3

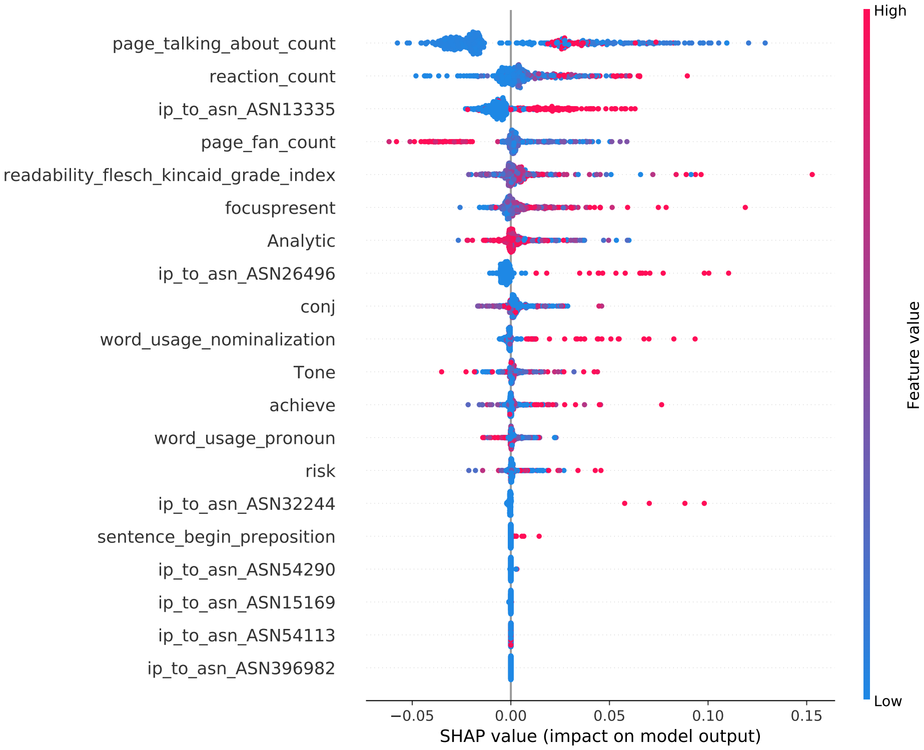


(b) Fold 4

Figure C.19: SHAP summaries for the highest AUC model in **cluster 2** for the US election dataset (Folds 3 and 4).

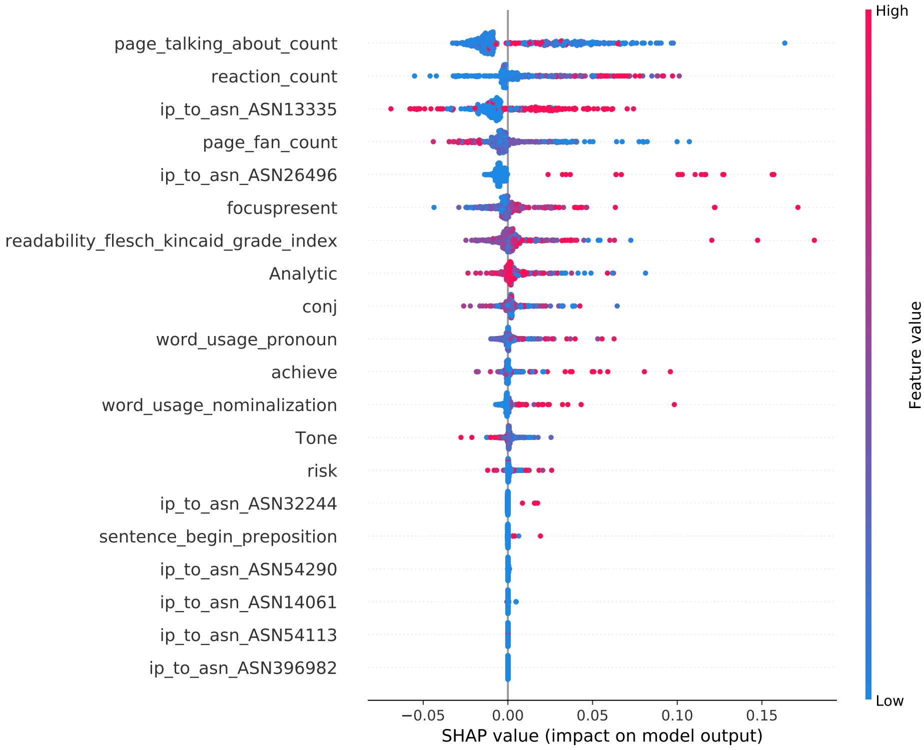


(a) Fold 1

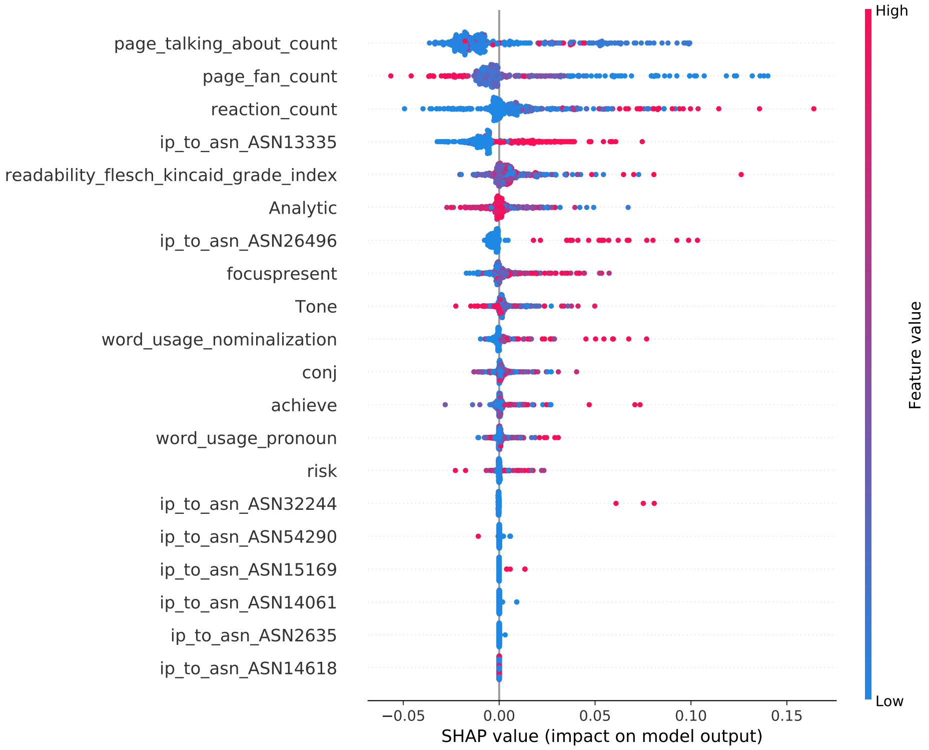


(b) Fold 2

Figure C.20: SHAP summaries for the highest AUC model in **cluster 3** for the US election dataset (Folds 1 and 2).

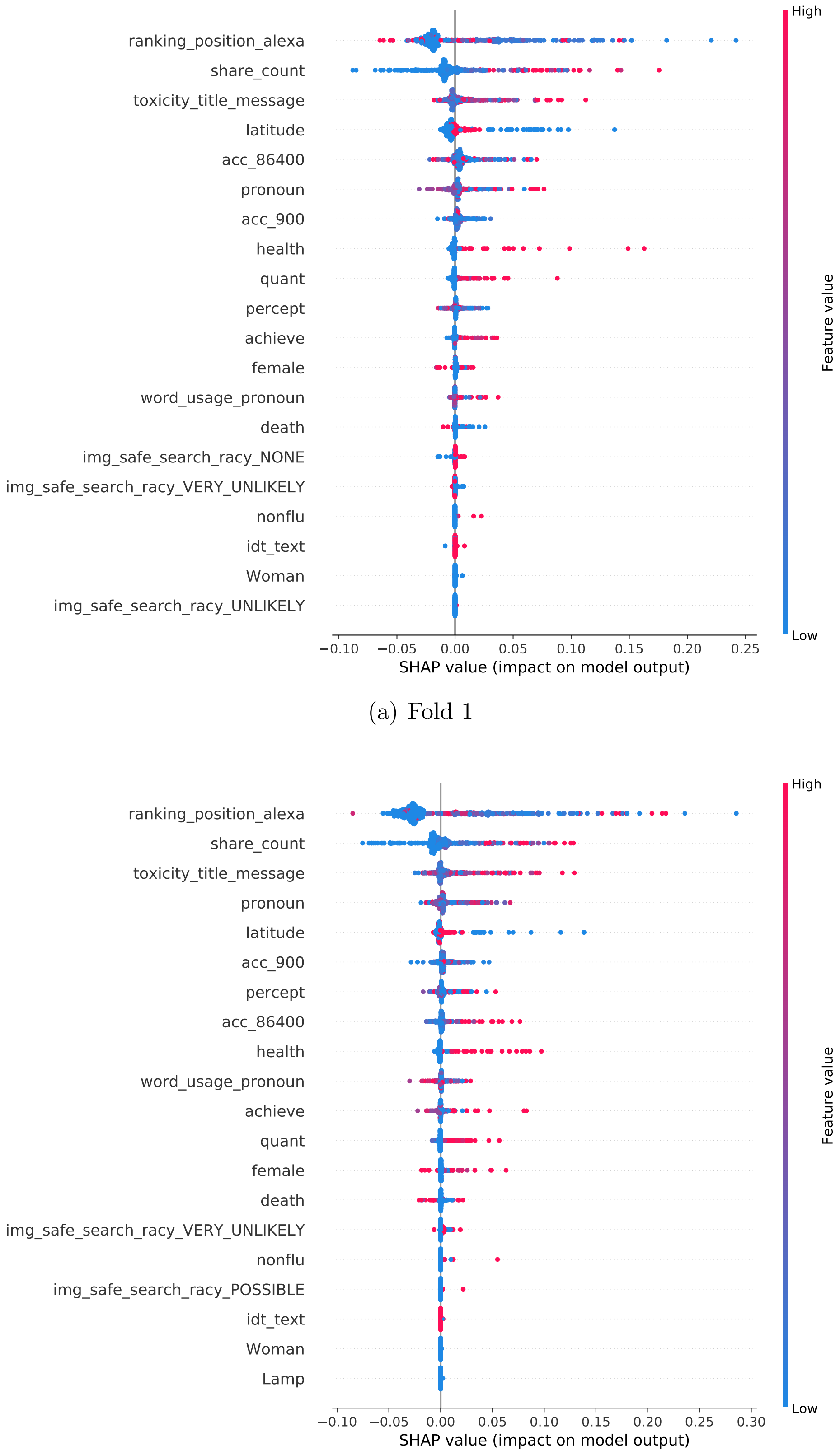


(a) Fold 3



(b) Fold 4

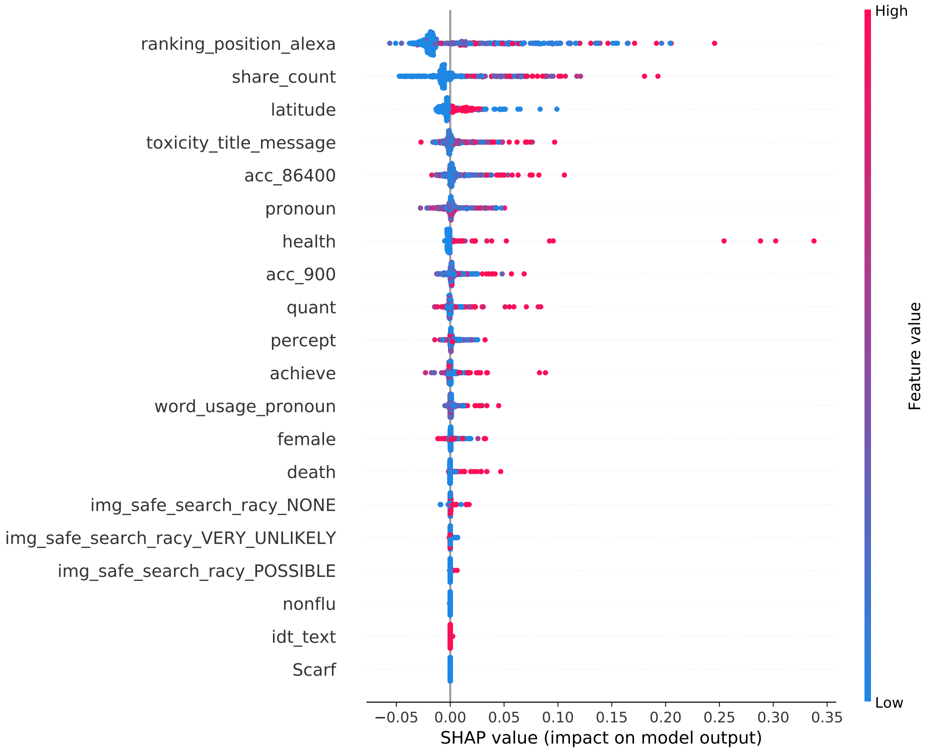
Figure C.21: SHAP summaries for the highest AUC model in **cluster 3** for the US election dataset (Folds 3 and 4).



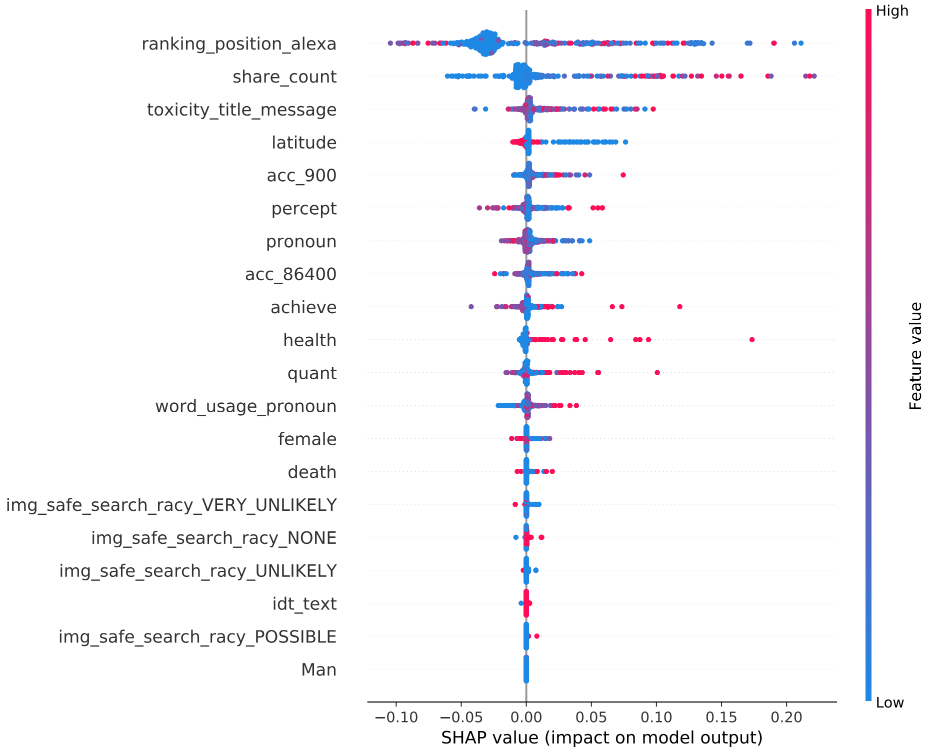
(a) Fold 1

(b) Fold 2

Figure C.22: SHAP summaries for the highest AUC model in **cluster 4** for the US election dataset (Folds 1 and 2).

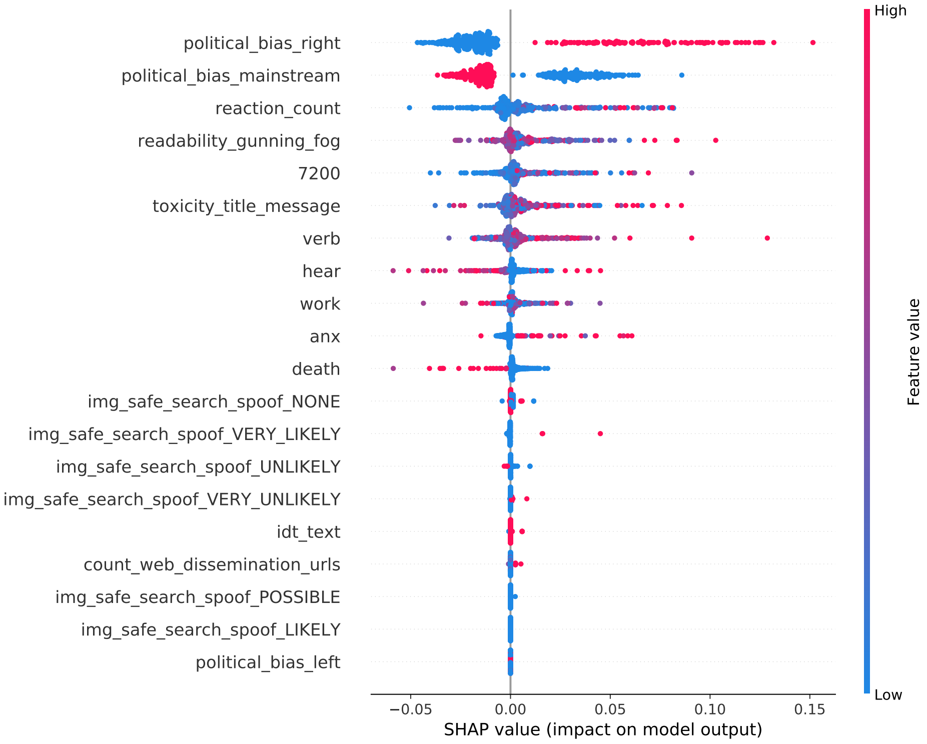


(a) Fold 3

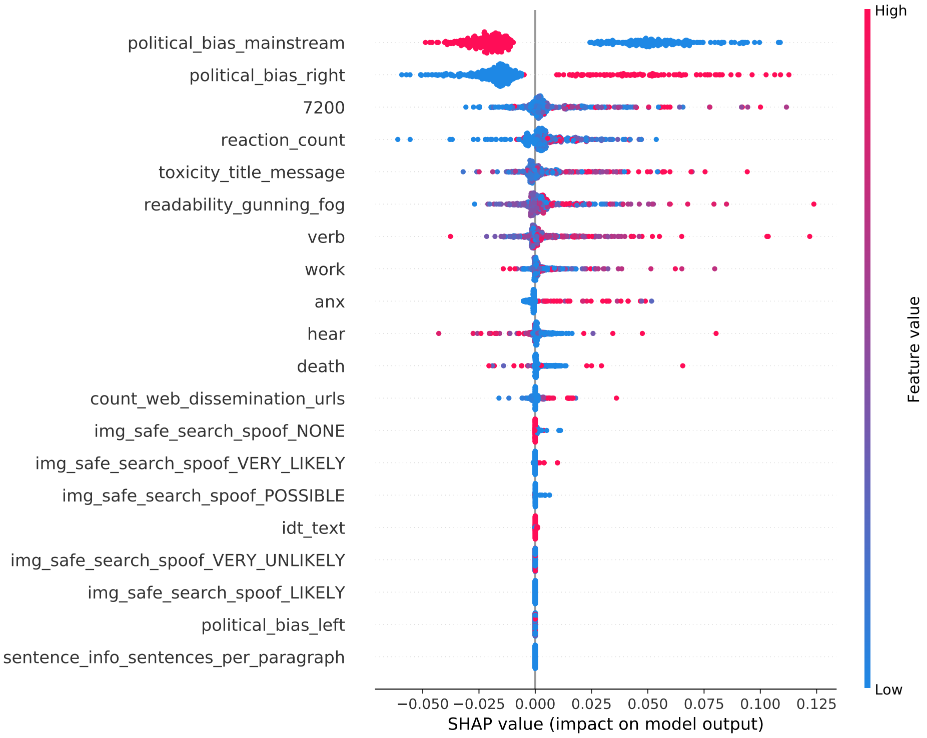


(b) Fold 4

Figure C.23: SHAP summaries for the highest AUC model in **cluster 4** for the US election dataset (Folds 3 and 4).

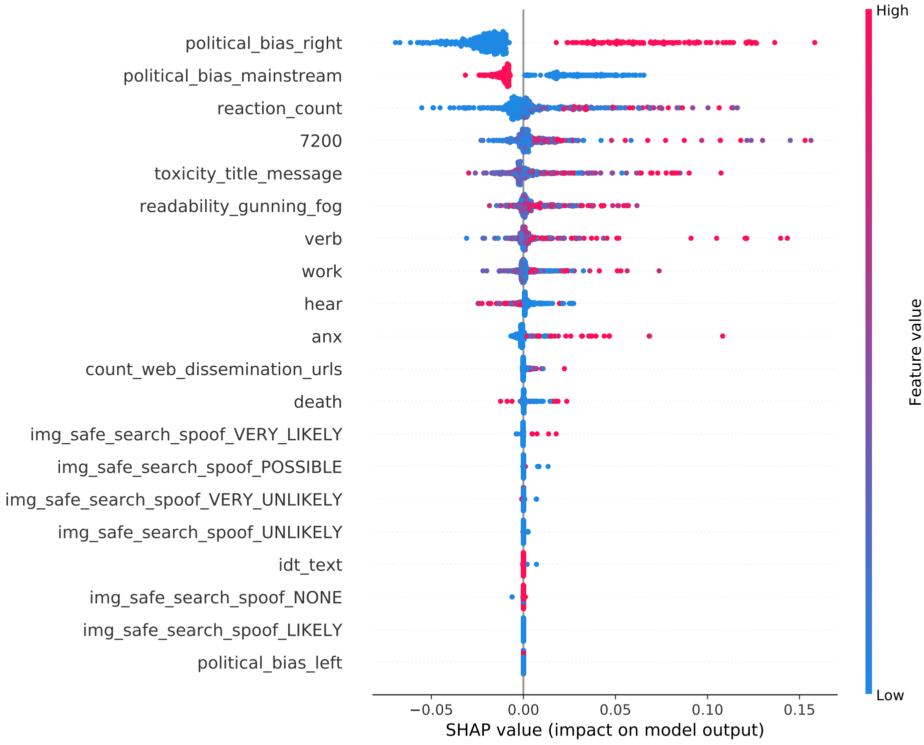


(a) Fold 1

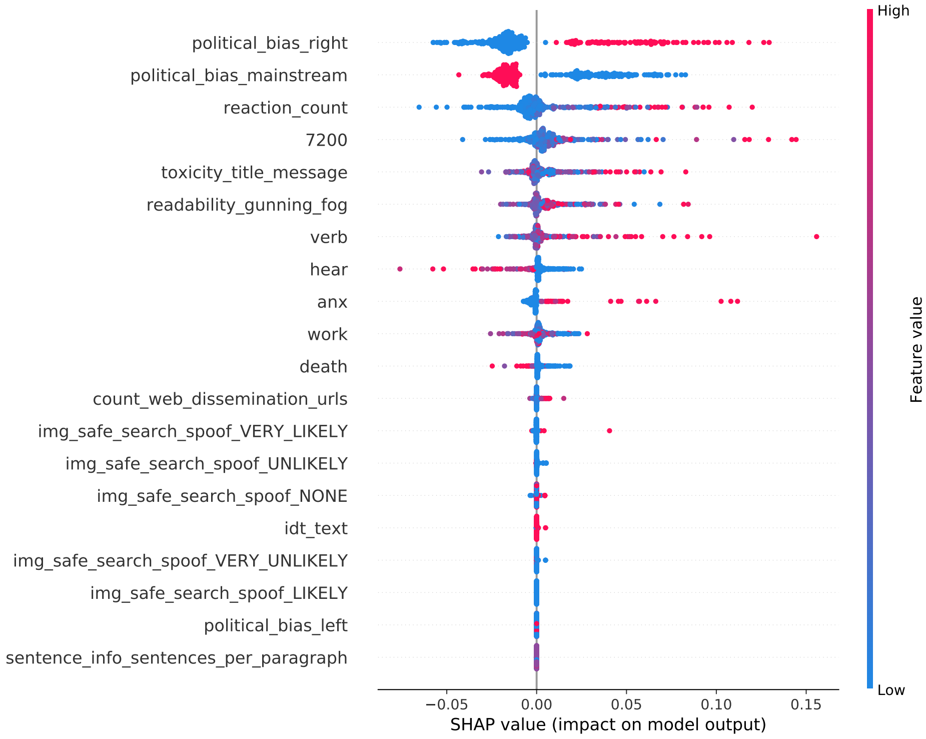


(b) Fold 2

Figure C.24: SHAP summaries for the highest AUC model in **cluster 5** for the US election dataset (Folds 1 and 2).

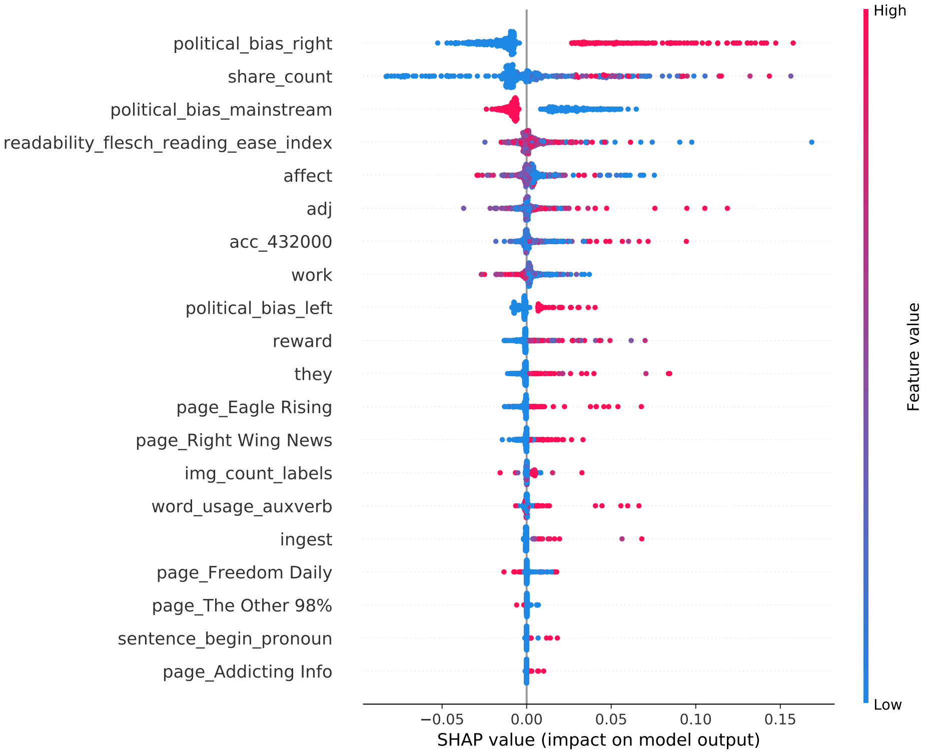


(a) Fold 3

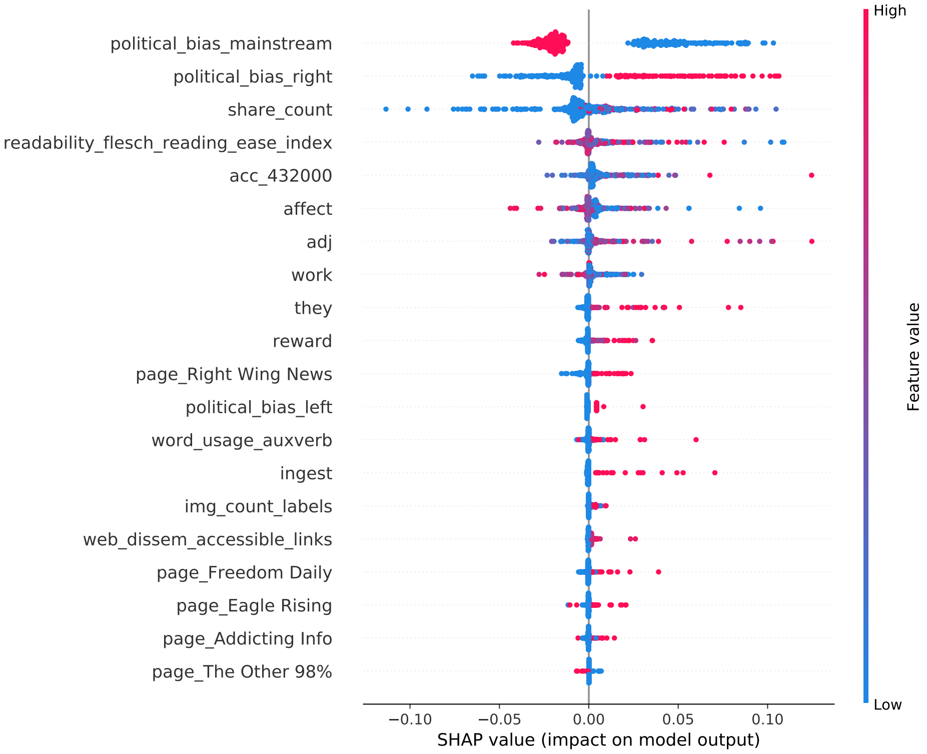


(b) Fold 4

Figure C.25: SHAP summaries for the highest AUC model in **cluster 5** for the US election dataset (Folds 3 and 4).

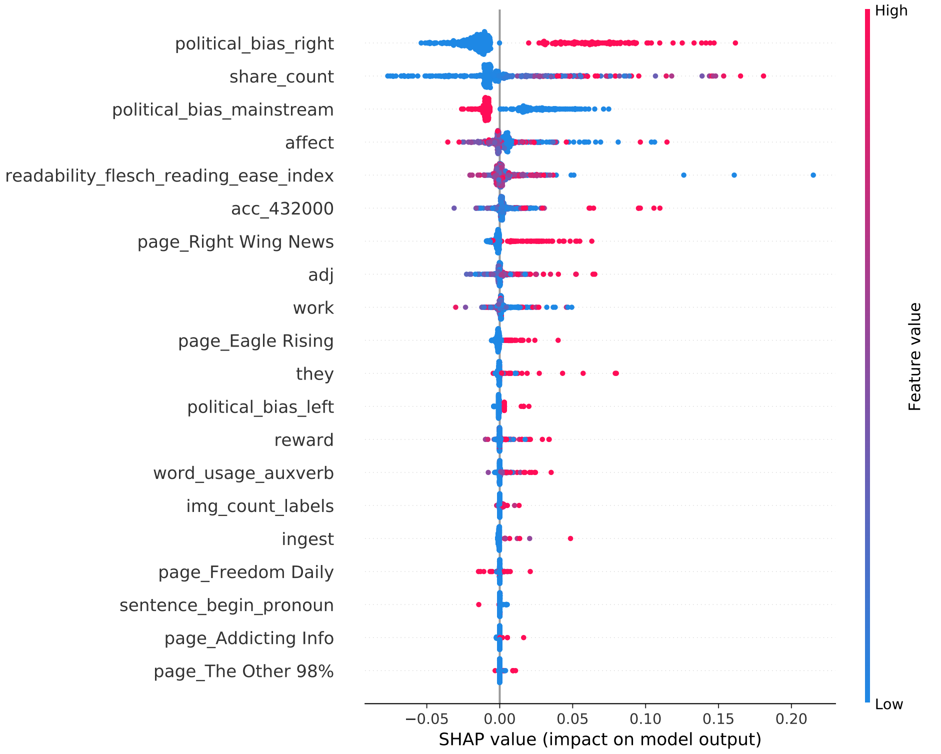


(a) Fold 1

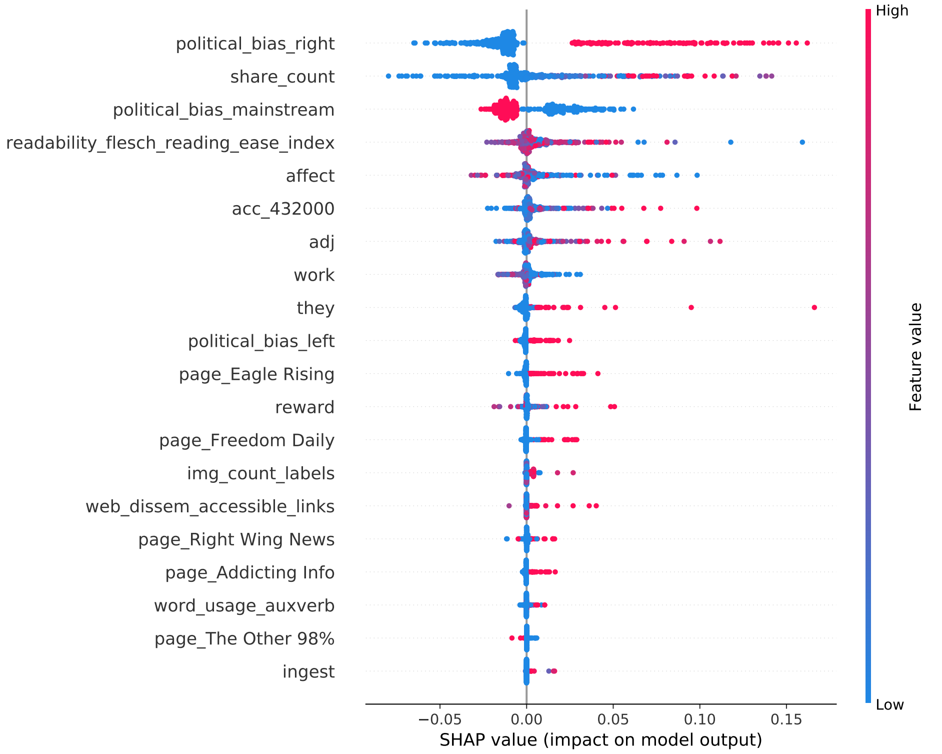


(b) Fold 2

Figure C.26: SHAP summaries for the highest AUC model in **cluster 6** for the US election dataset (Folds 1 and 2).

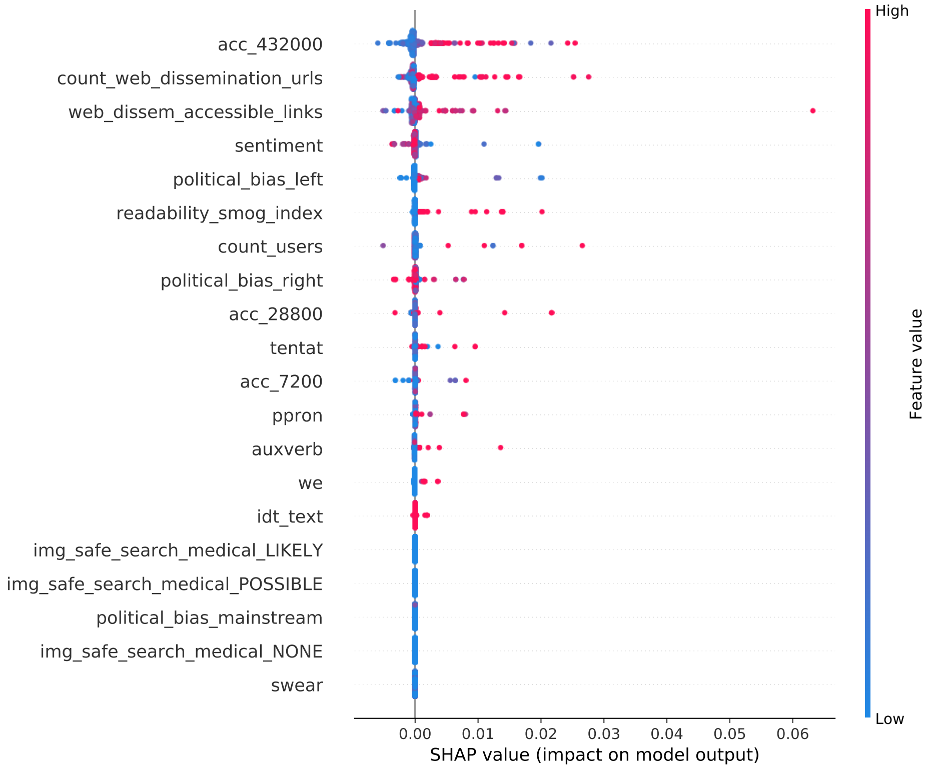


(a) Fold 3

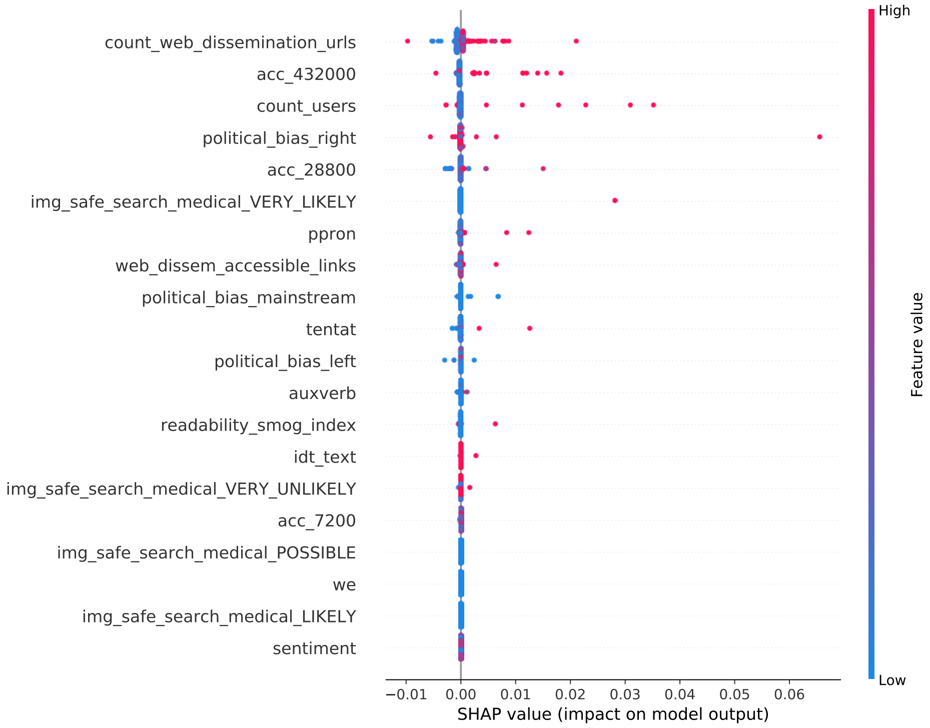


(b) Fold 4

Figure C.27: SHAP summaries for the highest AUC model in **cluster 6** for the US election dataset (Folds 3 and 4).

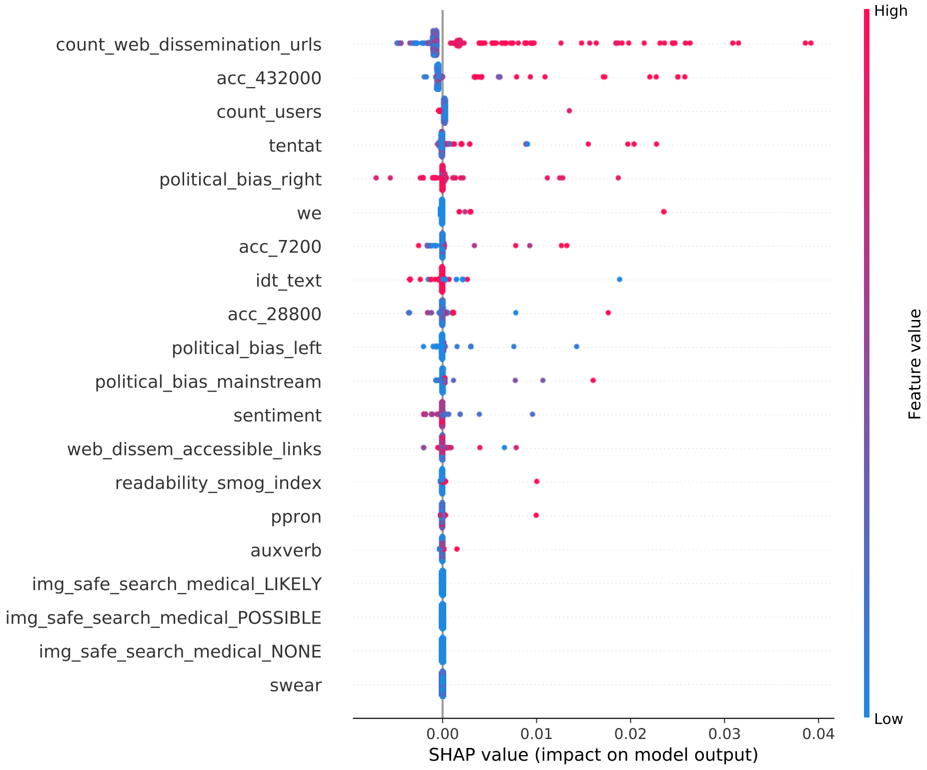


(a) Fold 1

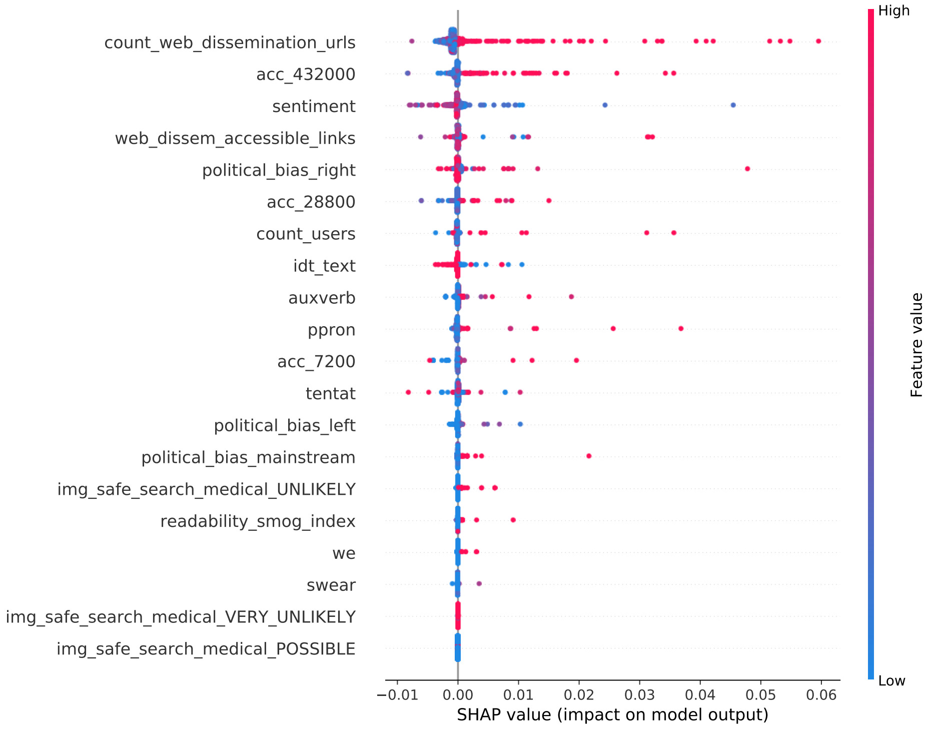


(b) Fold 2

Figure C.28: SHAP summaries for the highest AUC model in **cluster 1** for the Brazilian election dataset (Folds 1 and 2).

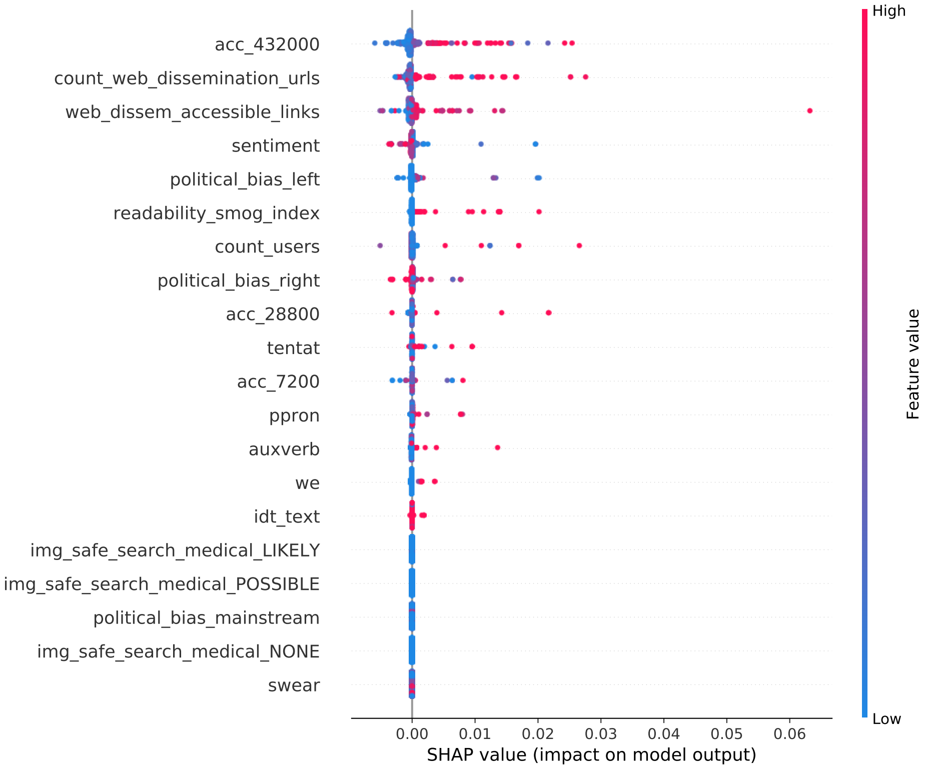


(a) Fold 3

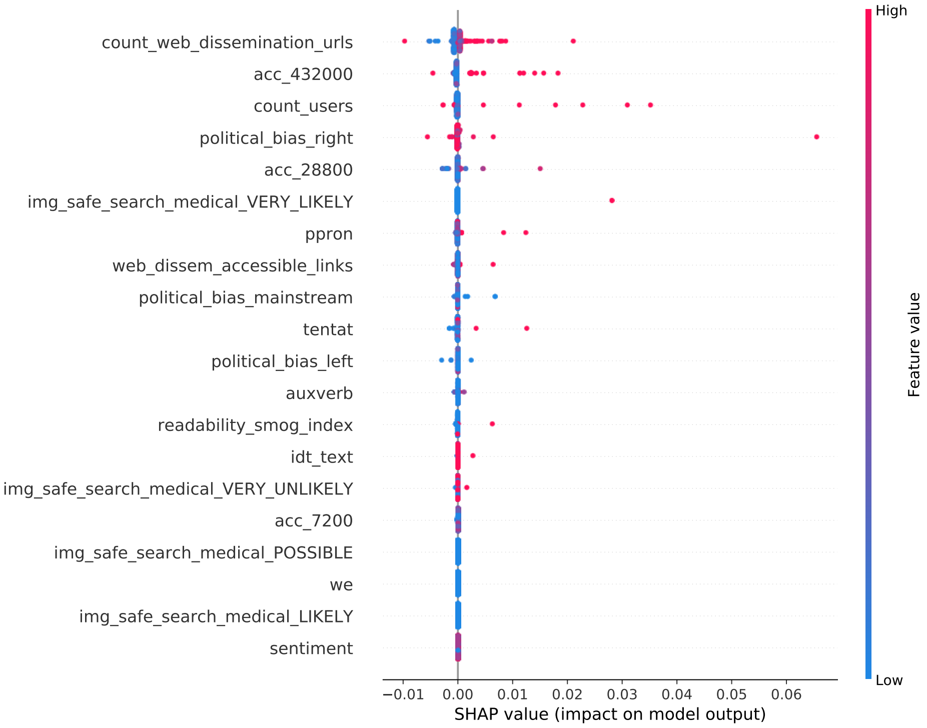


(b) Fold 4

Figure C.29: SHAP summaries for the highest AUC model in **cluster 1** for the Brazilian election dataset (Folds 3 and 4).

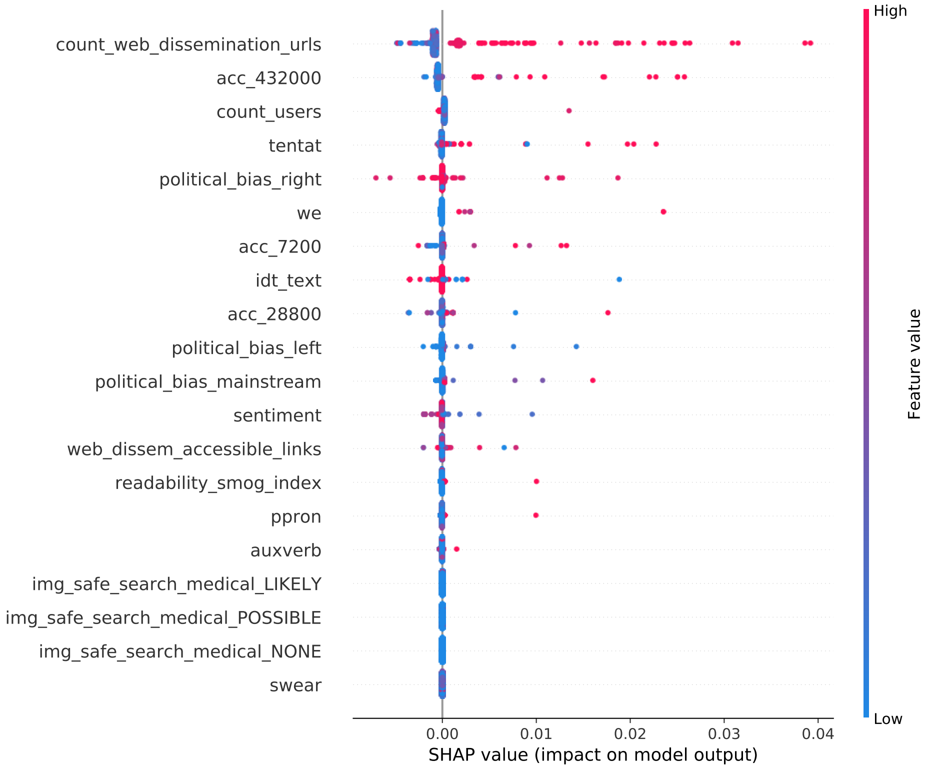


(a) Fold 1

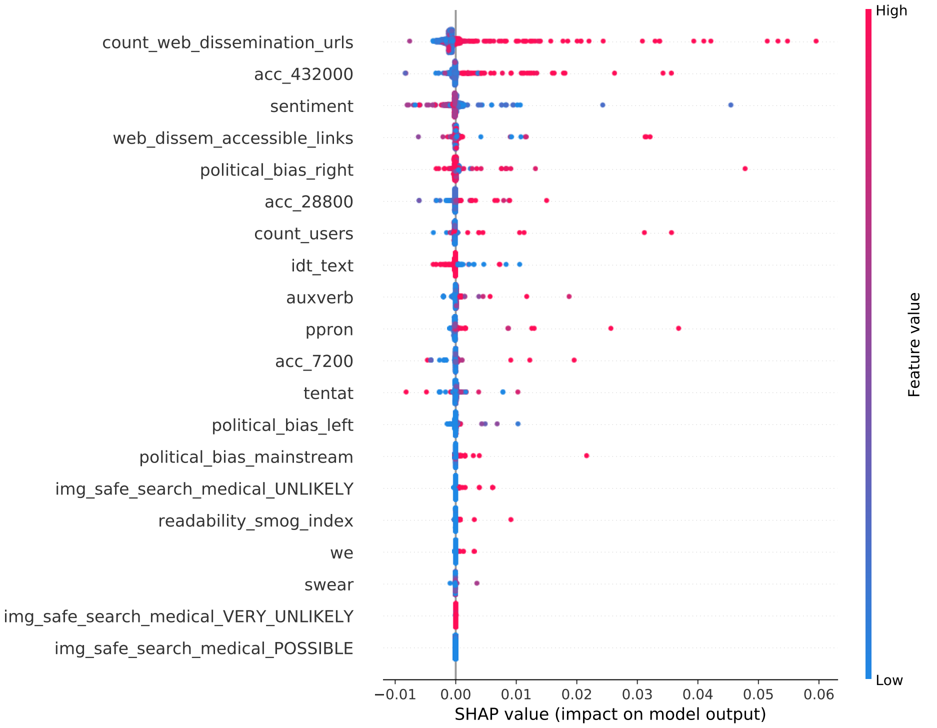


(b) Fold 2

Figure C.30: SHAP summaries for the highest AUC model in **cluster 2** for the Brazilian election dataset (Folds 1 and 2).

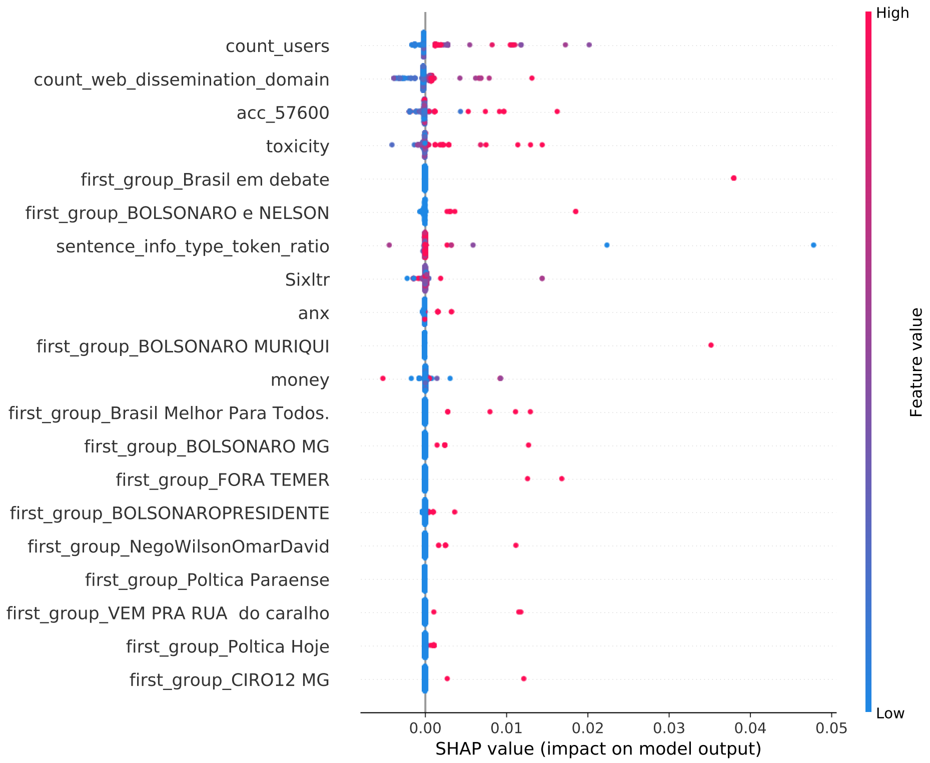


(a) Fold 3

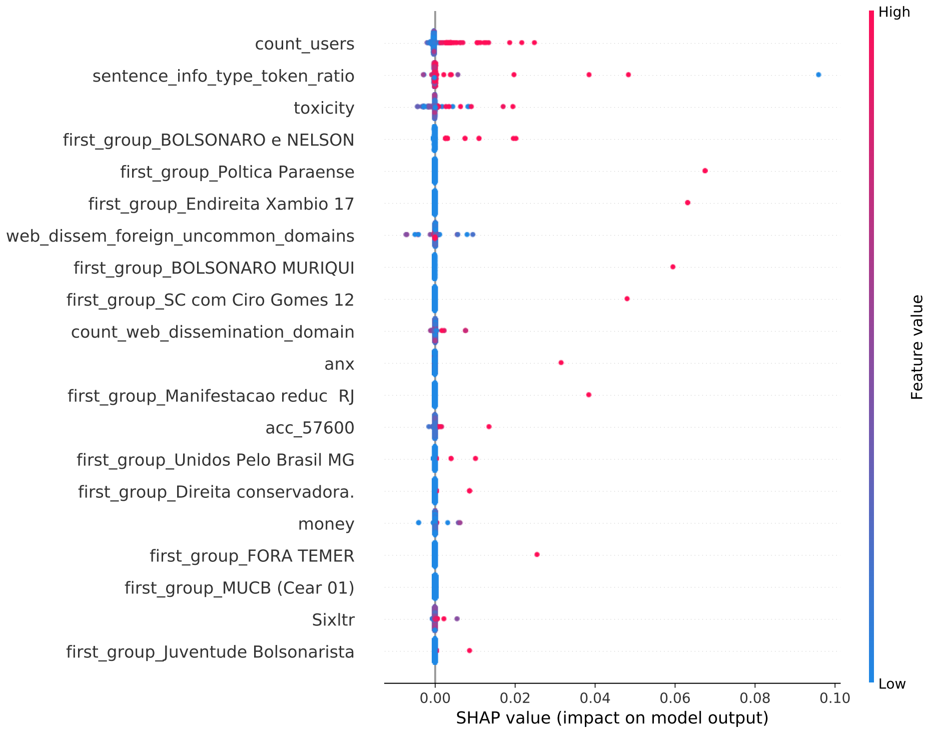


(b) Fold 4

Figure C.31: SHAP summaries for the highest AUC model in **cluster 2** for the Brazilian election dataset (Folds 3 and 4).

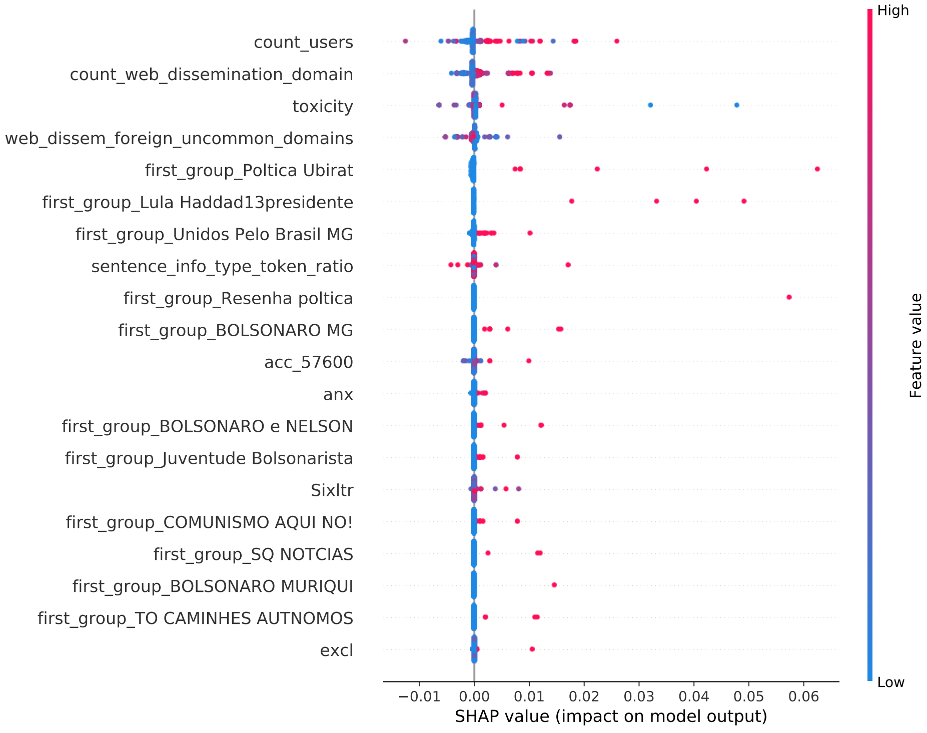


(a) Fold 1

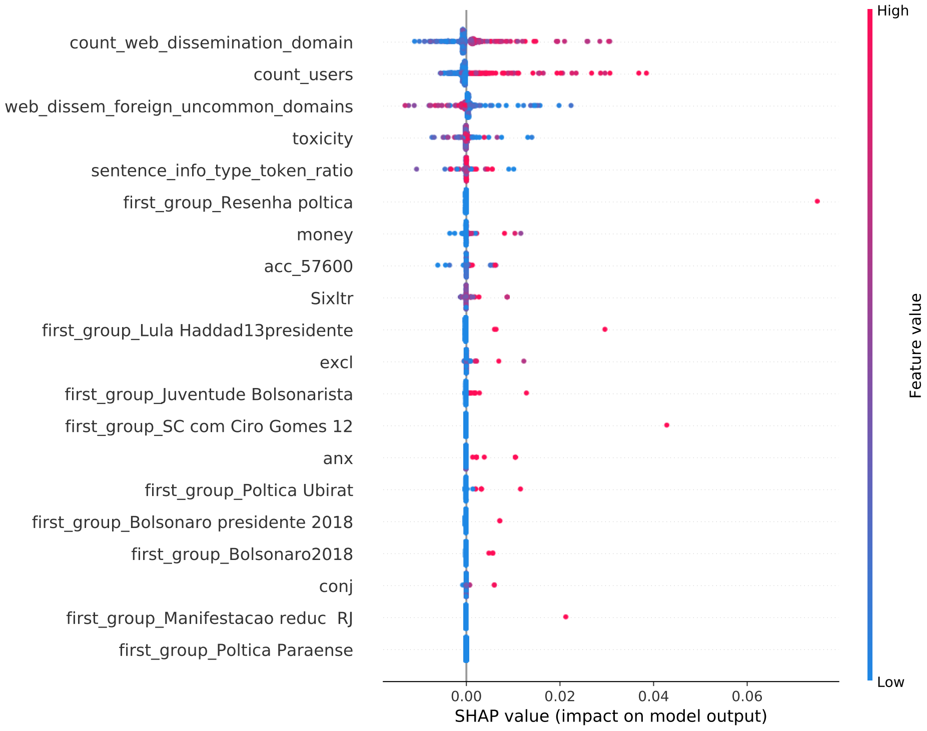


(b) Fold 2

Figure C.32: SHAP summaries for the highest AUC model in **cluster 3** for the Brazilian election dataset (Folds 1 and 2).

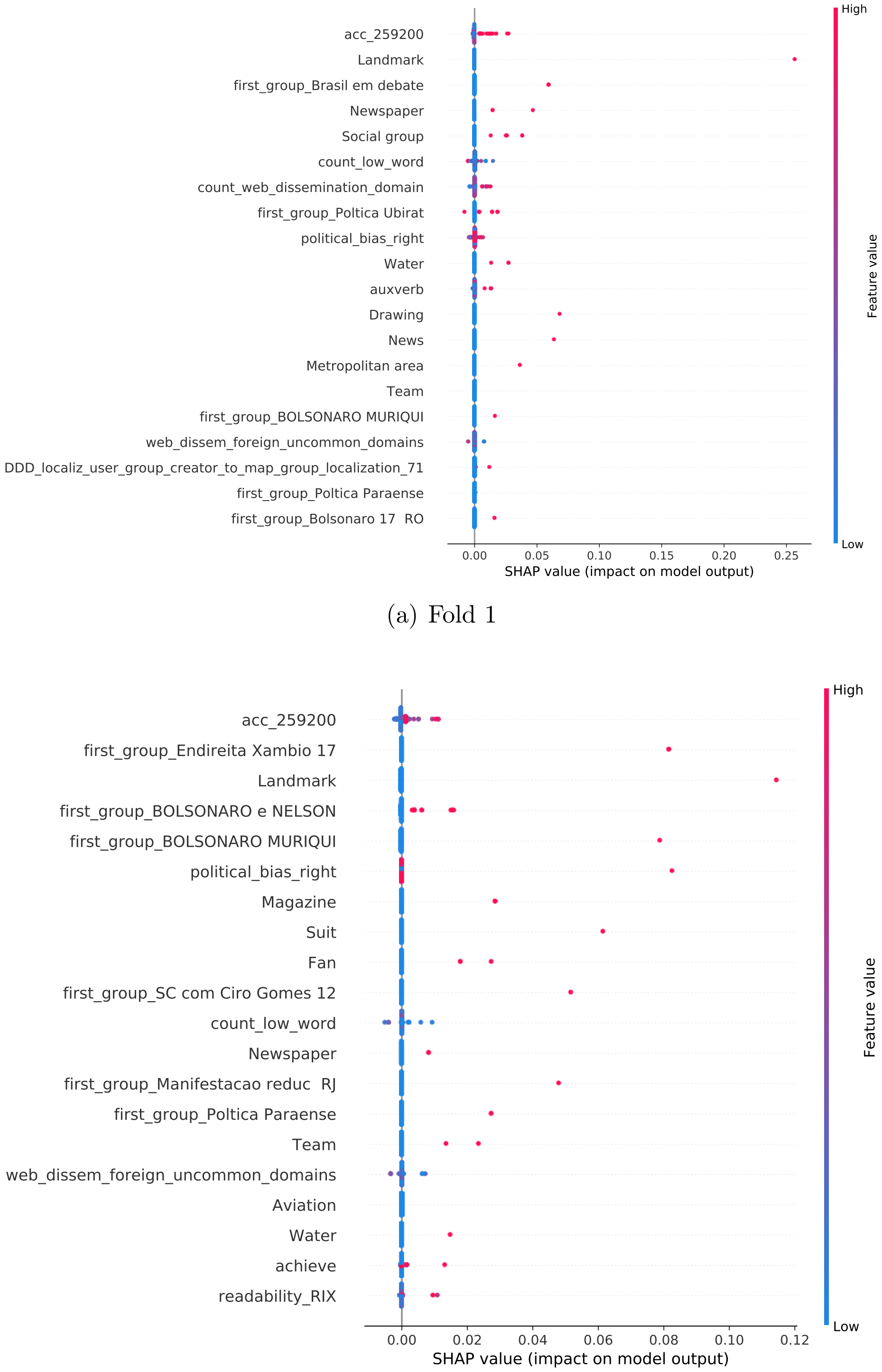


(a) Fold 3



(b) Fold 4

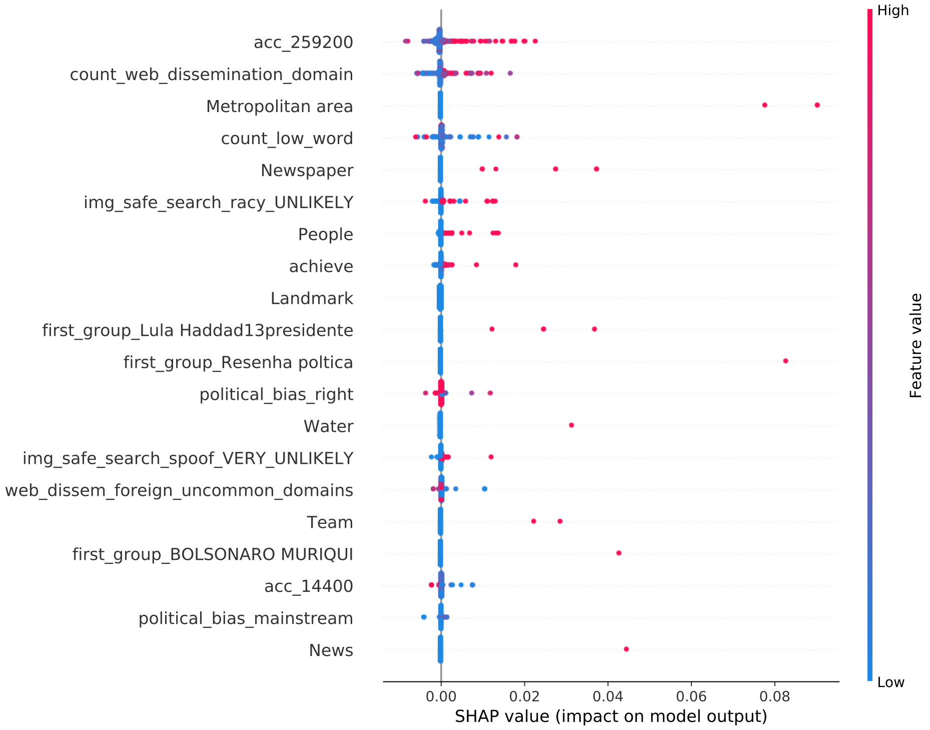
Figure C.33: SHAP summaries for the highest AUC model in **cluster 3** for the Brazilian election dataset (Folds 3 and 4).



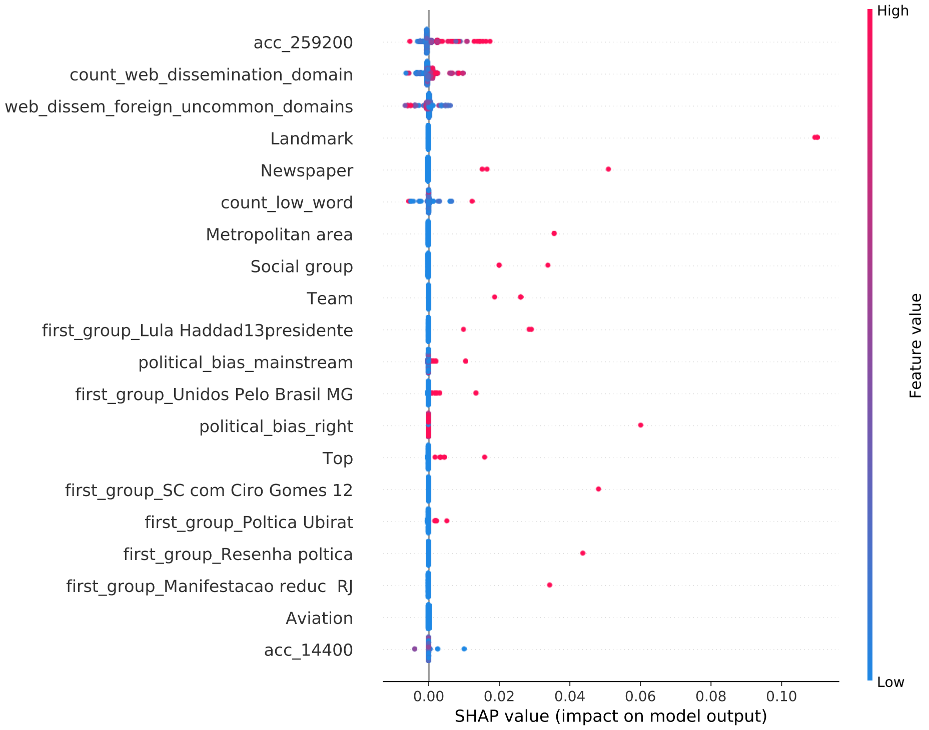
(a) Fold 1

(b) Fold 2

Figure C.34: SHAP summaries for the highest AUC model in **cluster 4** for the Brazilian election dataset (Folds 1 and 2).



(a) Fold 3



(b) Fold 4

Figure C.35: SHAP summaries for the highest AUC model in **cluster 4** for the Brazilian election dataset (Folds 3 and 4).