



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Fake News Classification in European Portuguese Language

João Filipe Carriço Rodrigues

Master in Integrated Business Intelligence Systems

Supervisors:

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Doctor Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

November, 2020



TECNOLOGIAS
E ARQUITETURA

Fake News Classification in European Portuguese Language

João Filipe Carriço Rodrigues

Master in Integrated Business Intelligence Systems

Supervisors:

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Doctor Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

November, 2020

Resumo

Um pouco por todo o mundo foram tomadas várias iniciativas para combater fake news. Muitos governos (França, Alemanha, Reino Unido e Espanha, por exemplo), à sua maneira, começaram a tomar medidas relativamente à responsabilidade legal para aqueles que fabricam ou propagam notícias falsas. Foram feitas algumas mudanças estruturais nos meios de comunicação sociais, a fim de avaliar as notícias em geral. Muitas equipas foram construídas inteiramente para combater fake news, mais especificamente, os denominados "fact-checkers". Essas equipas têm vindo a adotar diferentes tipos de técnicas para realizar as suas tarefas: desde o uso dos jornalistas para descobrir a verdade por detrás de uma declaração controversa, até aos cientistas de dados, que através de técnicas mais avançadas como as técnicas de Text Mining e métodos de classificação de Machine Learning, apoiam as decisões dos jornalistas. Muitas das entidades que visam manter ou aumentar a sua reputação, começaram a concentrar-se em elevados padrões de qualidade e informação fiável, o que levou à criação de departamentos oficiais e dedicados de verificação de factos. Na primeira parte deste trabalho, contextualizamos o Português Europeu no âmbito da deteção e classificação de notícias falsas, fazendo um levantamento do seu actual estado da arte. De seguida, apresentamos uma solução end-to-end que permite facilmente extrair e armazenar notícias portuguesas europeias previamente classificadas. Utilizando os dados extraídos aplicámos algumas das técnicas de Text Mining e de Machine Learning mais utilizadas, apresentadas na literatura, a fim de compreender e avaliar as possíveis limitações dessas técnicas, neste contexto em específico.

Palavras chave

Fake News, Portuguese European Language, Fact-checking, Web Scraping, Text Mining, NLP, Machine Learning, Deep Learning

Abstract

All over the world, many initiatives have been taken to fight fake news. Governments (e.g., France, Germany, United Kingdom and Spain), on their own way, started to take actions regarding legal accountability for those who manufacture or propagate fake news. Different media outlets have also taken plenty initiatives to deal with this phenomenon, such as the increase of the discipline, accuracy and transparency of publications made internally. Some structural changes have been made in those companies and in other entities in order to evaluate news in general. Many teams were built entirely to fight fake news, the so-called “fact-checkers”. Those teams have been adopting different types of techniques in order to do those tasks: from the typical use of journalists, to find out the true behind a controversial statement, to data-scientists, in order to apply forefront techniques such as text mining, and machine learning to support journalist’s decisions. Many of those entities, which aim to maintain or rise their reputation, started to focus on high standards of quality and reliable information, which led to the creation of official and dedicated departments of fact-checking. In the first part of this work, we contextualize European Portuguese language regarding fake news detection and classification, against the current state-of-the-art. Then, we present an end-to-end solution to easily extract and store previously classified European Portuguese news. We used the extracted data to apply some of the most used text minning and machine learning techniques, presented in the current state-of-the-art, in order to understand and evaluate possible limitations of those techniques, in this specific context.

Keywords

Fake News, Portuguese European Language, Fact-checking, Web Scraping, Text Minning, NLP, Machine Learning, Deep Learning

Acknowledgements

In first place I would like to thank my supervisors, Ricardo Ribeiro and Fernando Batista, for all the motivation and support throughout the whole work as well as all the knowledge they have passed, that was essencial to complete this work and critical in my professional future.

Secondly, I would like to acknowledge my parents, Maria Rodrigues and António Rodrigues, by helping and motivating me to accomplish all my objectives. Furthermore for never giving up even when the hardest times appear, which was essential during all my academic life.

Lastly, an acknowledgment to all of my closest friends, specially Maria Vilalobos, Kevin Ganhão, João Custódio, Manuel Tavares, Ana Barros, Paula Moreira and Rui Marques, where each one of them, in it's own way, was essential in the support and consequent ending of this dissertation.

Lisboa, November 2
João Filipe Carriço Rodrigues

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Goals and Research Questions	3
1.3	Document Structure	3
2	Related Work	5
2.1	Fake News	5
2.2	Datasets	7
2.2.1	Fake News Challenge (FNC-1)	7
2.2.2	Fake.Br Corpus	9
2.2.3	BuzzFace	11
2.2.4	WSDM Cup	12
2.3	Methods	14
2.4	Summary	16
3	FakePT	19
3.1	Fact-checkers	21
3.1.1	Polígrafo	22
3.1.2	Observador	23
3.1.3	Coronaverificado	24
3.2	Web Scraping	25
3.3	Database Import	26
3.4	Dataset Overview	26
3.4.1	Characteristics	27
3.4.2	Exploratory Data Analysis	28

4 Fake News Detection	35
4.1 Data Pre-Processment	35
4.1.1 Data Quality Assurance	36
4.1.2 Feature Extraction	41
4.2 Data Selection	41
4.2.1 Feature Selection	42
4.2.2 News Selection	42
4.2.3 Data Split	42
4.3 Data Classification	43
5 Results	45
5.1 Evaluation Metrics	45
5.2 Experiments	46
5.2.1 Phase 1	46
5.2.2 Phase 2	47
5.3 Discussion	49
6 Conclusions and Future Work	53
6.1 Conclusions	53
6.2 Contributions	54
6.3 Future Work	55
Bibliography	57

List of Figures

3.1	IFCN Poynter Principles	22
3.2	Polígrafo Tags	23
3.3	Observador Tags	24
3.4	Coronaverificado Tags	25
3.5	News Example Polígrafo	28
3.6	Classified News by Fact-checker and by Month	29
3.7	Polígrafo Category Values Distribution	31
3.8	Observador Category Values Distribution	32
3.9	Coronaverificado Category Values Distribution	32
3.10	Polígrafo Source Values Distribution	33
3.11	Observador Source Values Distribution	33
3.12	Coronaverificado Source Values Distribution	34
3.13	All News Tag Values Distribution	34
4.1	Fake News Classification Pipeline	35
4.2	Tags Classification Assessment	40

List of Tables

2.1	Records by Label (FNC-1)	8
2.2	Pre-processing tasks and Textual Representations by Author (FNC-1)	8
2.3	Records by Label (Fake.Br Corpus)	9
2.4	Pre-processing tasks and Textual Representations by Author (Fake.Br Corpus)	10
2.5	Records by Label (BuzzFace)	11
2.6	Pre-processing and Textual Representations by Author (BuzzFace)	11
2.7	Records by Label (WSDM Cup)	12
2.8	Pre-processing and Textual Representations by Author (WSDM Cup)	13
2.9	Summary Table of Authors Approaches 1	16
2.10	Summary Table of Authors Approaches 2	17
3.1	Database Table Fields	27
3.2	FakePT Dimensions	27
3.3	News by Fact-checker	28
3.4	Number of Words by Dimension	31
4.1	Category Cleaning Steps by Fact-checker	37
4.2	Source Cleaning Steps by Fact-checker	39
4.3	Feature Selection	42
4.4	Classifiers	43
5.1	Experiments Phase 1	48
5.2	Experiments Phase 2	50

Introduction



“Freedom depends upon citizens who are able to make a distinction between what is true and what they want to hear. Authoritarianism arrives not because people say that they want it, but because they lose the ability to distinguish between facts and desires.”

- Timothy Snyder

The way in which each of us, regular consumers of contents in the environment that surround us, bridges ignorance and knowledge has been changing over time. This bridge, the channel responsible for making available and disseminating “common interest” content, has been suffering changes in its form, content and perception of reliability, from the consumer’s perspective. Contrary to the period prior to the beginning of the Internet, these interventions, that moved according to the political, economic, social and scientific context of each society, are now at the mercy of a new context that has been gaining strength in recent years – technology. Since the beginning of our existence, until the early 2003, humanity has generated 5 Exabytes of data [1]. Today, that same volume is produced in only two days. In parallel with this fact, the access points to every kind of information also grew, both for information and misinformation. Today, as we live in a world where the surging of the aforementioned data is dramatic, where the struggle for audiences on traditional media increases and new forms of information are now found in uncontrollable proportions, and where the thoroughness in the management and proliferation of information is declining, it is urgent to provide ourselves a critical and attentive eye to fight the avalanche of disinformation to which we are exposed every day. In the last decade, more traditional information channels such as newspapers and television have been forced to give space, and consequently power, to a new giant phenomenon that has been conquering the market – the social media. This migration of content consumption is essentially due to the popularity that certain social platforms, such as Facebook and Twitter, have started to gain in society [7]. With the emergence of social media, both positive and negative aspects have shown up in terms of impact for its users. On one hand, social networks have brought to life a tool that, due to its regular use in conjunction with its massive popularity, allowed not only an easy way to search for others, but also a huge ease in the almost instantaneous proliferation of news. Hence, news with transverse interest to the entire population, such as the reporting of events in times of crisis, can be obtained almost in real time, either through official news

channels or by any user who uses social platforms. Despite this positive side, social media have also seemingly harmed our society in a variety of ways and fields of interest [29]. In the traditional media field, the way that these large entities reached their listeners had to be rethought and reformulated, since there was a major shift in the interest in consumption of news by their target audience towards social media. In fact, nowadays it is quite easy for a user without any contractual affiliation to an audiovisual entity to achieve more views, in a specific content shared only by himself, than some contents presented on FOX News, CNN, and New York Times [3]. Media outlets, in order to avoid completely losing the race for the attention of their target audience, were forced to emphasize and focus on the number of views / clicks of their publications at the prejudice of their content [7]. Ethics, integrity and accountability have been transformed into sensationalism, views number and, ultimately, greed. With this, the perfect environment for the appearance of fake news is settled. News with a willful lack of attention to source confirmation, fake news, misleading news, rumors and especially click bait news, which the only goal is to attract the user to a content that seems to be relevant and interesting, but, after a quick glance at said content, it ends up being far below the user's expectations [2].

However, despite the fact that fake news are recently growing in the most traditional media outlets, it was not here that they have taken such proportions for the first time. In 2016, after the elections in the United States of America that resulted in the victory of Donald Trump, it was immediately possible to realize, in a clear way, that great consequences come from the proliferation of fake news on a large scale. Many were those who addressed this topic and concluded that most of the most debated fake news preceding the election favored Donald Trump over Hillary Clinton. Furthermore, it is unanimous that Donald Trump would never win the elections without the influence of these fake news [3].

The exploration of tasks such as the detection and classification of "Fake News" is quite recent. Due to the enormous media exposure in which the subject has been involved, essentially after the North American elections in 2016 where, allegedly, Russian influence jeopardized the outcome of the results [11], a boom of contributions and initiatives began to take form. Together, the global community started to develop a common sense of urgency to address the problem and started, as a whole, searching for different solutions and approaches [7]. Detection and classification tasks deal with unstructured textual information using NLP (Natural Language Processing) techniques. Different techniques and methodologies provide quite different results. A large number of these tools can be applied within the scope of this dissertation.

1.1 Motivation

Nowadays it is increasingly difficult to trust the information we find at our disposal. We live in an era in which we have never had easier access to information, but it is increasingly diffi-

cult to know how to classify it as reliable or not. The competition between traditional media and emerging media is increasing, causing a greater temptation in the production and proliferation of cheap and sensationalist news. Traditional media are less and less dependent on themselves. It is increasingly notorious the enthrall of many media companies towards the interests of great economic, political and social powers, interests that often overlap with the interests of the common citizen and even with the rule of law. The competition between these powers is so big that slander environments between people, products and companies are easily promoted without looking at their practical and moral consequences [16]. The work to contain these initiatives is also precarious and it is regrettable to see that the judicial means and the entities that have the regular task of preventing violations of the journalistic code of ethics, are quite ineffective. Thus, in a world in which the creation of fake news grows visibly and where their impacts can reach tragic proportions in society, it is increasingly important to seek to create solutions for their identification, classification and mitigation. This is where this dissertation is inserted.

1.2 Goals and Research Questions

As main objective we propose the creation of one of the first datasets (FakePT) in European Portuguese (PT-PT) that allows the exploration and analysis of the phenomenon of fake news in the national context. We also propose the subsequent application of the best approaches of text processing (basic and NLP techniques) and also classification (traditional machine learning and deep learning) found in literature. In order to successfully achieve our goals, we propose to answer the following research questions:

- How to create a Fake News dataset?
- Which pre-processing techniques, textual representation tasks and classification methods are most relevant in the Fake News scope?
- What are the limitations that the Portuguese language appoint to success in the classification of Fake News?

1.3 Document Structure

This document is divided into four chapters and is structured as follows:

Chapter 1 is dedicated to the Introduction. It begins with a brief historical introduction to the theme of disinformation, alerting to the possible social, economic and political consequences over time. Then, and in more depth, the need for this work as a tool to fight fake news is discussed. It ends with the definition of the objectives and the research questions that we propose to answer.

Chapter 2 is dedicated to Literature Review. This chapter is divided into three sections. The first section is dedicated to the detailed analysis of the fake news concept. It begins with a historical introduction of several mainstream episodes of fake news over time and ends with a comparison of the many definitions of fake news according to the point of view of various authors, from an individual perspective to the national and international entities perspective. The second section is dedicated to exploring the most popular datasets used in fake news. A characterization of the dataset is made, from its purpose to its constitution. Finally, a detailed analysis is made of some works that use datasets as a working and exploration tool, ending with a comparison of all the different approaches. In the third and last section, an analysis is made of the classification techniques identified in the literature, both machine learning and deep learning, and a comparison of methods and their performances is made.

Chapter 3 is dedicated to FakePT dataset creation description. This chapter will cover all the work developed from the news extraction to the creation and exploration of the dataset (FakePT). Also, an introduction is made to the concept of “Fact-checking”, and all the fact-checkers that are used in this work are then described.

Chapter 4 is dedicated to the Fake News Classification Pipeline. An exhaustive exploration will be made on the best pre-processing techniques, textual representation and classification methods.

Chapter 5 is dedicated to the Results. Here, all the evaluation metrics used in our tests are described and, also, all the experiments and respective discussion over the experiments results is presented.

Chapter 6 is dedicated to Conclusions and Future Work. This chapter is dedicated to the evaluation of all the work itself, from what went well to the various limitations encountered. An analysis and discussion of the results is done. Finally, a reflection is made on the contribution of our work and an identification of possible future challenges is made.

Related Work

2

“When you can admit that you don’t know, you are more likely to ask the questions that will enable you to learn”

- Richard Saul Wurman

In this section, we will begin by conducting an analysis of the different definitions of "Fake News". Subsequently, a survey and its respective analysis will be presented, regarding the most used and recommended datasets by the academic and business community in an attempt to find a resolution to this type of problems and concerns. For each one of them, a summary is made with a brief introduction to its structure and composition, a mention to the authors who contributed most to its exploration and the most used features in each of the datasets. Finally, the last point of this section deals with the different methods/classifiers used in this type of cases, associating, to each one, the datasets in which they were used and the respective authors.

2.1 Fake News

The idea behind the "Fake News" concept is not a novelty. In fact, if we go back a few years, to the time when the Internet did not even exist, and despite the terms not being the exact same, the idea of misinformation and disinformation was already circulating in society and in the traditional media of these times. There are many historical examples that support the aforementioned statement. As early as 1835, for instance, the first major hoax manufactured by the New York Sun newspaper appeared, where countless articles reporting the discovery of life on the moon were published [3]. Already in 2006, another major fabrication of false news appeared in Belgium, where a Belgian television station reported that the Flemish parliament had declared independence from Belgium [3]. At this time, although the market was not as fragmented as it is today, and the fact there were very few media outlets, the power that these entities had at their disposal was already unanimously recognized. Information is power, and the greater the power, the greater the appetite for promiscuous interests and consequent corruption [8].

As for the term Fake News itself, its definition began to gain popularity in the period during and after the North American elections of 2016, which culminated in the election of Donald Trump as president of the United States [3]. Also, alongside this term, two other are now commonly used to describe this phenomenon: misinformation and disinformation. However, these words are often incorrectly used for the same purpose, while they mean two different things. Both are used to refer to any spread piece of content that is false or misleading, however, there is a key difference – the intent behind each content. “Disinformation”, in contrast to “Misinformation”, is a term used for describing any false or misleading content that is intentionally spread while knowing about the content’s lack of truth.

Many papers have emerged after Donald Trump’s election, and plenty suggested different definitions for the term “Fake News”. The most consensual definition used by most scholars emphasizes the importance of intention and verification: fake news are any and all new news that are proven intentionally false [3]. Therefore, any news that, through different sources, can be disproved, proving to be categorically false, or any news through which the author has the clear intention of misleading, are considered as fake news.

Other authors look at the concept of fake news from different perspectives. In [8], a definition is given regarding three different aspects: publications based on manufactured content; publications inserted in the context of large fraudulent and defamatory campaigns; and humorous publications. The work done by [7] also mentions three different aspects – humorous content, the need for verifiability, and a new perspective presented as “malicious content”. According to these authors, it is necessary to consider the humorist content as misleading, since although it is directed to a public that recognizes the author’s humorist intention at the outset, there is also a large portion of the community that fails to correctly identify it. Along with this definition comes the concept of “malicious content”, which represents the definition of intention, as mentioned above (disinformation).

The European Commission, a political independent entity with numerous goals, such as the creation of legislation, policies and action programs that cross the interests of the whole European Union, suggested, in 2018, the creation of a group of highly specialized experts on the subject of fake news. This group had a mission – to ensure that the democratic process of the 2019 European elections ran without any kind of interference from both misinformation and disinformation contents. At the time, this group defined the operational concept of Fake News as “all information which is proven to be false or misleading and which is created, presented and disseminated in order to obtain economic advantage or to deliberately mislead the public, and which is likely to cause public harm” [11]. This was the first definition to address the importance of public harm and economic advantages. With these, public harm arises and it is likely the most blatant and alarming consequence of the fake news phenomenon. The authors define it as all the threats to democratic political processes and public goods such as health, environment and security. On the other hand, we also have one of the causes of the immense growth of Fake News contents and consequent

public harm – economic interests. Economic lobbies, entities who usually place their interests above the common human being, are one of the most important reasons for the large investment in fake news. This, combined with political interests as well, causes a great deal of pressure and influence on traditional media, coming from multinational companies, the state itself and magnanimous entities.

Finally, in the National context, the definition of fake news by the ERC, the Regulatory Authority for the Media in Portugal, appears. In 2019, the ERC conducted a study entitled "Disinformation - European and National Context", which deals with this subject in great detail [11]. Fake News are defined by this entity in a very similar way to the one given by the European Commission, but it adds a new perspective which should be taken into account in this context. ERC mentions that a news story, in its definition, can never be false, but that the contents of the narratives that are inserted in it, can be false or misleading. This means, according to the author, that labelling any news as a "Fake New" might be a little abusive, semantically speaking, and misleading.

It was also from this study that the subject in the next chapter, FakePT, took shape.

2.2 Datasets

This section is dedicated to the detailed analysis of the most used datasets in the literature in the context of fake news. For each one an overview is made, specifying when, by whom and for what purpose it was created. An analysis of each one's content is also made, such as the number of records, the different labels and the variables that define it. Finally, a comparative analysis of different works is presented, defining which made use of each of the datasets, the approach taken by each one regarding the basic textual processing performed, the NLP tasks performed, and the textual representations implemented, which will serve as input features to the machine learning algorithms presented in the next section.

2.2.1 Fake News Challenge (FNC-1)

The creation of this dataset took place in a challenge called "Fake News Challenge", in 2017, and counted on the joint effort of 100 volunteers from the academic and business fields [20]. The goal of this challenge was to find new methods and approaches that would be useful to present solutions to fight fake news. This dataset contains a set of news written in English and about 50,000 associations of statements with news. Each statement is associated with a particular news item and also with a label. The dataset is rather unbalanced in the sense that it has four different labels with a scatter sample distribution between them 2.1.

This dataset has a different nature than the one expected in a typical Fake News problem. Often this problem is thought and addressed for the classification of titles and news bodies from a dichotomous perspective (usually called "truth labeling"), i.e., true or false.

Label	Records	Percentage
Unrelated	36,545	73.1%
Disagree	8,909	17.8%
Agree	3,678	7.4%
Discuss	840	1.7%

Table 2.1: Records by Label (FNC-1)

The purpose of this challenge is different: it is about trying to understand what is the "posture/relationship" of one statement before another, or a set of others (news/text body). Thus, the objective is to classify an affirmation using one of the previously mentioned labels: "Discuss" means that the body text neither confirms nor denies the statement; "Unrelated" means that there is no relationship between the body text and the statement; "Disagree" means that the body text does not agree with the statement; "Agree" means that the statement and the body text are related and agree with each other. According to Pomerleau and Rao [20], this approach is not a substitute for truth labeling, but a means of supplementing it. The choice was made based on talks with journalists and fact-checkers where both parties mentioned that it is quite difficult to make truth labeling classification. Both mentioned that they would rather have a semi-automatic solution to assist them in their work than a fully automatic solution whose performance would be far below expectations.

Many studies were based on this challenge and the respective dataset, and today this is one of the most used and explored datasets by all fake news researchers. Thus, each author has tried to approach it in his own way, using different text processing approaches and different types of representations of the text. These different representations form the features that will serve as input to the later machine learning and deep learning tasks 2.2.

Author	Pre-processing		Textual Representation
	Basic processing	NLP	
[14]	Regular expressions	Lemmatisation Stanford NER	Glove Embeddings
[6]	Label mapping	Sentiment Analysis	Word-Embeddings
	Tokenization		Google News CNN
	Stemming		Number of n-grams
	N-grams generation		TF-IDF
			word2vec
			SVD
			Number of positive and negative words

Table 2.2: Pre-processing tasks and Textual Representations by Author (FNC-1)

In Kotonya and Toni [14], the authors started by using regular expressions to eliminate all unwanted links. Following that, they made use of lemmatization – the process

of grouping together the inflected forms of a word as a single item. The Stanford Entity Name Recognizer [13] was also used to replace entities such as names, organizations, and locations. Finally, pre-trained GloVe embeddings models [18] were used to represent all words in global semantic vectors. Baird, Sibley, and Pan [6] describe the work that led them to win the first place in the competition. The "Solat In the Swan" team chose to use the combination of two classification approaches, and to do so they needed to create two sets of features. In order to develop both sets it was necessary, first of all, to perform a pre-processing step. In this phase, the tokenization of both titles and news were made and stemming was applied to each of the tokens. Finally, the various unigrams, bigrams and trigrams were generated from the list of tokens. Once the pre-processing was carried out, the first textual representation was created. For this, pre-trained Google News vectors were used, both for the titles and the news themselves. More traditional features were applied in the second set: n-grams counts; TF-IDF; SVD-based; word2vec and sentiment features.

2.2.2 Fake.Br Corpus

For the Portuguese language, Fake.Br Corpus is the only dataset we found in this context. According to Monteiro et al. [16], this is the first fake news dataset with Brazilian Portuguese news and also the first dataset in the Portuguese language. The dataset resulted from a joint effort of researchers and analysts who gathered and classified news manually. The dataset contains 7200 news items and is divided in a balanced way regarding the number of records that are associated with the different labels – Table 2.3.

Label	Records	Percentage
True	3,600	50%
False	3,600	50%

Table 2.3: Records by Label (Fake.Br Corpus)

The authors defined a time span of two years (January 2016 to January 2018) and only collected news that were inserted in it. Some news, however, reference other news in previous time spaces. Other relevant information, such as the author of the news, the date of publication, the number of views and comments, were kept. The news were manually tagged and, for each false one, the authors used a semi-automatic process to find true news that could prove the tag on the corresponding "false" ones. The false news were manually extracted from four different newspapers. After this process, through web scrapping, 40,000 real news were taken from other sites, based on the most frequent words, verbs and names that were in each of the fake news previously taken. Then, a lexicon similarity measure, cosine, was applied to determine which were the true news that were closest to the fake news. After this process was completed, having already a smaller number of news

Author	Pre-Processment		Textual Representation
	Basic Processment	NLP	
[16]	Stopwords Removal	Stemming	Bag of Words
	Punctuation Removal		
	-	POS Tagging	Number of each Part of Speech takes place
	-	Enriched Lexicon	Number of semantic classes
	Words Removal	-	Pausality (number of punctuation characters)
	Extraction of Some Words (can, might)	POS Tagging	Expressiveness (sum of adjectives and adverbs over the sum of nouns and verbs)
	-	-	Incertainty (number of modal verbs)
	Extraction of Some Words (can, might)	POS Tagging	Non-Immediacy (number of first and second pronouns)
[4]	Stopwords Removal	Stemming	Word-Embeddings
		Lemmatisation	
		Chi-square	

Table 2.4: Pre-processing tasks and Textual Representations by Author (Fake.Br Corpus)

items, they made the manual selection of the news items based on their actual degree of similarity.

Once again, the work on this dataset explored different pre-processing procedures and different types of text representations – Table 2.4.

In [16], the authors use various approaches to the representation of the news text. Apart from the most common forms of textual representation (bag of words, term count, etc.), the authors also explore the use of linguistic features that may also be interesting, such as pausality, uncertainty, expressiveness, non-immediacy, and number of semantic classes.

Andrade [4] presents two alternative approaches to the work of [16]. First, mentions the use of the chi-square method as a means of selecting the most relevant terms that resulted from his pre-processing task. According to his work, this method is an added value for textual processing tasks in the Portuguese language. Finally, he states that textual representations based on word embeddings have shown better results than classical representations based on term-to-document matrices.

2.2.3 BuzzFace

This dataset was developed from a set of news items published by the media company “BuzzFeed” after the 2016 North American elections and was manually tagged by journalists [24]. This dataset was also enriched with information from Facebook, such as comments, number of shares and reactions to the news published by the company on its website.

The dataset has a total of 2,282 entries. Each entry line corresponds to a share on Facebook from an official media outlet’s page. Each post can have one of four possible classifications: "no factual content", "mostly true", "mostly false" and "mixture of true and false" 2.5.

Label	Records	Percentage
True	1,665	73%
Non Factual Content	274	12%
Mixture of True and False	251	11%
Mostly False	91	4%

Table 2.5: Records by Label (BuzzFace)

This dataset is also of great importance since it has features related to the "context" of the news, i.e., number of shares, reactions and comments to the news post. This is the only dataset contemplated in this study that contains these types of features, which also means that it is the only one that does not depend directly on the intrinsic content of the news.

Author	Pre-processing		Textual Representation
	Basic processing	NLP	
[22]	-	-	Bag of words
	-	POS Tagging	Number of pronouns, verbs, adverbs, hashtags, punctuation.
	By “Linguistic Inquiry and Word Count”	By “Linguistic Inquiry and Word Count”	Psycholinguistic features (detection of biased and persuasive language)
	By “Google’s API”	By “Google’s API”	Semantic features (toxicity)
	By “Text Blop’s API”	By “Text Blop’s API”	Features subjectivity (subjectivity and feeling)

Table 2.6: Pre-processing and Textual Representations by Author (BuzzFace)

In [22], the authors start by ensuring a greater balance of the dataset, making the distribution of the records (for each label) more uniform. Thus, only two labels were considered: true and false, all cases with the label "mostly true" became "true"; the cases with

the label "no factual content" were removed; and the records with the two remaining labels were converted to the label "false". The authors then apply and detail a set of features appropriate to the problem in question, in which context-related features are also included. The features are thus divided into three main groups: (1) features extracted from the news content, (2) features extracted from the source of the news and (3) features extracted from the environment 2.6.

2.2.4 WSDM Cup

This dataset is one of the most recent in the literature and was developed with the purpose of being the object of study for those who participated in the challenge of the international conference WSDM (Web Search and Data Mining), WSDM Cup, which took place in 2019, organized by the ACM (Association for Computing Machinery).

The dataset in question was developed by ByteDance, a Chinese Internet and technology company, which owns an online news platform. One of the greater challenges faced by ByteDance is the fight against fake news. To this end, the company has created a database to collect all kinds of fake news, so that all news can be properly verified regarding their veracity before being presented in the platform [15].

The dataset was originated from the database mentioned above, counting with a total of 360,767 records. Each record has in its constitution a title of a false news A, a title of a news B (news to be classified) and its label (Agreed, Disagreed or Unrelated). The objective here is to try to understand if the news item B addresses the same subject and agrees with the title A (agreed), which makes the news item B false; to try to understand if the title B does not agree with the news item A (disagreed), making the news item B true; or to identify that the news B has no relation whatsoever with news item A (unrelated). The dataset has news titles in two different languages, Chinese and English. Its configuration follows a weight of 75% for training and 25% for testing, and presents the label distribution shown in Table 2.7:

Label	Records	Percentage
Unrelated	246,764	68.4%
Agreed	104,622	29%
Disagreed	9,379	2.6%

Table 2.7: Records by Label (WSDM Cup)

Table 2.8 shows some of the works using this dataset.

Author	Pre-processing		Textual Representation
	Basic processing	NLP	
[15]	Dataset augmentation	-	Set of 25 pre-trained BERT's.
	Stopwords removal		
[19]	Dataset increase	N-grams	Text Based
	Text to lowercase		Statistics
	Add spaces between punctuations		Graph Based
	Tokenization		KNN (BERT's)

Table 2.8: Pre-processing and Textual Representations by Author (WSDM Cup)

"Travel", the team that came second in the competition, developed an approach that achieved a "weighted accuracy score" of 0.88 [15]. Given the nature of the problem, the authors have chosen to attack the problem using NLI (Natural Language Inference) techniques. NLI is a subarea of NLP and its objective is to recognize textual implications – RTE (Recognizing Textual Entailment) – that is, in this case, from news B which is given as false, to be able to infer a hypothesis (which label characterizes news A) from a textual premise through the semantic similarities between them. Regarding the creation of the features that will feed the machine learning algorithms, the authors in question decided on three main steps. First, since the dataset was not properly balanced regarding the distribution of records per label, and in order to avoid an over-fit, an increase of quantity of the data was made through the transitivity of semantics. That is, if title A is related to title B and if title B is related to title C, then A and C are related. Subsequently, all stop words were removed, both in Chinese and English news. Finally, to address the problem of text representation, the authors chose to use BERT [10], a pre-trained linguistic model created by Google that has recently been gaining popularity.

In [19], the challenge's winning team ("IM"), suggests a pre-processing approach and textual representation similar to that used by the "Travel" team, but with some nuances. As far as pre-processing is concerned, the author suggests also separating text scores (by placing spaces) and the tokenization of all titles. In the textual representation, an ensemble of features is suggested for input regarding future classification algorithms. The first set of features are "Text Based". Here all textual features are covered, such as the generation of n-grams of words and characters. After their generation, distance measurements are applied to pairs of titles, such as cosine, euclidean, city-block, jaccard, or simple addition and subtraction. The second set of features are "Statistics". This is where word counts are present, as well as stop words, tokens, characters, or a simple comparison of the textual size of the title pairs. The third set of features are the "Graph Based". In this case, the texts of the titles are represented as graph networks. The objective of these networks

is to make a representation of each title (a node of the graph) and of each pair of titles. Through metrics, such as minimum or maximum distance between nodes and news pairs, assumptions can be made about their relation. Finally, the features "KNN" represent BERT embeddings with reduced dimensionality.

2.3 Methods

This section is dedicated to the analysis of the most commonly used methods in fake news detection. There are several ways to approach this topic according to the literature, however, most approaches cast this as a classification problem. In a classification problem the aim is to be able to associate a label, for example true or false, with small (in the case of a title, for example) or large (in the case of a news' body text, for example) textual portions. In order to respond to this task, most of the research body dedicated to this subject implements machine learning and deep learning techniques [7].

Within the classification problem, authors employ different methods depending on the features they have at their disposal. Typically, most of the approaches focus on using features extracted from the news content itself. However, other sets of features, such as information related to the source or context of the news [22], present another type of detail that may enrich the analysis. Regarding fake news classification strategies, a combination of methods is typically used, the so-called "ensemble models". Since in most works that ensemble methods are used the evaluation metrics refer to the set and not to each method itself, it is relevant to make a comparative analysis not only between methods, but also between standalone methods and ensemble methods.

There are several methods of machine learning that have been applied in the task of fake news detection. Of the most recent and best performing methods, there are three that typically stand out from the more traditional methods (e.g., KNN, Naive Bayes (NB), Decision Trees (DT), etc). The first method is the SVM. The SVM (Support Vector Machine) is a discriminative classifier formally defined by a hyper plane of separation [7]. This method has been used in several fake news tasks. In two of the four datasets mentioned in the previous section, there are authors who propose the use of the SVM in isolation [16, 22].

In [16], in a study using the dataset "Fake.Br Corpus", the authors used the SVM with only content features. Among several combinations of textual representations, the best performance obtained was a F1-score of 0.89. In [22], in a study using the dataset "BuzzFace", the authors made use of all kinds of features and in all of them applied an SVM. In this work, the performance was a F1-score of 0.76, a performance that was below other methods also applied, such as Random Forest (0.81 F1-score) and Gradient Boosting (0.81 F1-score). An interesting approach using SVM [27] is "Graph-Kernel-Based SVM". This variation of SVM is used to identify rumors using propagation structures and content features. In this study, the authors reported an accuracy of 0.91.

Another method that has been gaining prominence over more traditional methods is the "Gradient Boosting" approach. Gradient Boosting is a meta algorithm based on decision trees, and it is used to reduce biased predictions. Catboost and LightGBM, for instance, are versions of Gradient Boost that have been gaining popularity in recent times due to their advantages of fast processing and high prediction performance [19]. These types of algorithms typically appear in stance detection problems, but can also appear in truth labeling problems [14, 19, 22]. In [14], in a specific stance detection problem, an accuracy of 0.83 was achieved for this method alone. Although attaining a good performance, gradient boosting ended up behind two other classification methods also tested: LSTM (Long Short-Term Memory) and BiLSTM (Bidirectional Long Short-Term Memory) neural networks models. In [22], in a specific truth labeling problem, gradient boosting was also used achieving the best performance (0.81 F1-score) against other algorithms: SVM, Naive Bayes, and Random Forest.

Finally, deep learning models have brought great advances in several areas of Artificial Intelligence, such as image identification, speech recognition, and textual processing [19]. In this topic, the most commonly used neural networks are CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks). In [4], a study on the dataset "Fake.Br" using neural networks, more specifically a CNN, achieved a performance of 0.91 accuracy which translated in a very successful result for a truth labeling problem. In [14], the authors using LSTM and BiLSTM neural networks managed to obtain a 0.92 and 0.93 accuracy, respectively, in this work regarding stance detection task.

As previously mentioned, ensemble methods have been a very successful approach to fake news detection. Typically combining deep learning and traditional machine learning techniques, these approaches achieved better results, in most cases. In the two public challenges that are described in this document, WSDM Cup and Fake News Challenge, the winning teams made use of an ensemble method. In the Fake News Challenge 2016 (FNC-1), the authors who came first [6], after several attempts to apply methods individually, concluded that the best performance was achieved by combining the methods they were exploring. Thus, the best performing approach was a combination of CNNs and Gradient-Boosting Decision Trees methods. This approach had an average weighted score of 82.02%. In [15, 19], works that finished first and second in the "WSDM Cup 2019" competition, the response to the problem also included a set of methods. The second placed team [15] used a total of 6 methods (3 SVM's, 1 Naive Bayes, 1 KNN and 1 Logistic Regression), obtaining an average accuracy score of 88.15%. The first team [19] chose to combine 28 methods (18 Neural Network Models, 9 Tree Based Models and 1 Logistic Regression), resulting in an average accuracy score of 88.28%.

2.4 Summary

Tables 5.1 and 5.2 summarize all the authors pre-processing and classifying approaches described, by each dataset used.

Author	Pre-Processment		Textual Representation	Methods	Dataset
	Basic Processment	NLP			
[22]	-	-	Bag of words	Gradient Boosting SVM Naive Bayes Random Forest	BuzzFace
	-	POS Tagging	Number of pronouns, verbs, adverbs, hashtags, punctuation.		
	By "Linguistic Inquiry and Word Count"	By "Linguistic Inquiry and Word Count"	Psycholinguistic features (detection of biased and persuasive language)		
	By "Google's API"	By "Google's API"	Semantic features (toxicity)		
	By "Text Blop's API"	By "Text Blop's API"	Features subjectivity (subjectivity and feeling)		
[15]	Dataset augmentation	-	Set of 25 pre-trained BERT's.	SVM Naive Bayes KNN Logistic Regression	WSDM Cup
	Stopwords removal				
	Tokenization				
	Data cleaning				
[19]	Dataset increase	N-grams	Text Based	NNM Tree Based Model Logistic Regression	
	Text to lowercase		Statistics		
	Add spaces between punctuations		Graph Based		
	Tokenization		KNN (BERT's)		
[27]	-	Content Features	-	Graph-Kernel-Based SVM	Rumors Dataset

Table 2.9: Summary Table of Authors Approaches | 1

Author	Pre-Processment		Textual Representation	Methods	Dataset
	Basic Processment	NLP			
[16]	Stopwords Removal	Stemming	Bag of Words	SVM	Fake.Br Corpus
	Punctuation Removal				
	-	POS Tagging	Number of each Part of Speech takes place		
	-	Enriched Lexicon	Number of semantic classes		
	Words Removal	-	Pausality (number of punctuation characters)		
	Extraction of Some Words (can, might)	POS Tagging	Expressiveness (sum of adjectives and adverbs over the sum of nouns and verbs)		
	-	-	Incertainty (number of modal verbs)		
-	POS Tagging	Non-Immediacy (number of first and second pronouns)			
[4]	Stopwords Removal	Stemming	Word-Embeddings	CNN	
		Lemmatisation			
		Chi-square			
[14]	Regular expressions	Lemmatisation	Glove Embeddings	LSTM	
		Standford NER		BiLSTM	
[6]	Label mapping Tokenization Stemming N-grams generation	Sentiment Analysis	Word-Embeddings	CNN Gradient Boosting	FNC-1
			Google News CNN		
			Number of n-grams		
			TF-IDF		
			word2vec		
			SVD		
			Number of positive and negative words		

Table 2.10: Summary Table of Authors Approaches | 2

“The point of modern propaganda isn’t only to misinform or push an agenda. It is to exhaust your critical thinking, to annihilate truth.”

- Garry Kasparov

The main objective of this work, as it was mentioned in the Introduction, is to study the Fake News phenomenon in European Portuguese, with the help of Text Mining techniques. Right now, as we demonstrated in the state of the art chapter, there are several datasets available in the literature to study Fake News. However, none of them have in its constitution news from Portugal, in European Portuguese (PT-PT). So, the previous point raises two important questions to be considered and understood in order to successfully study this phenomenon.

1. Does the language have any impact in Text Mining techniques?
2. What data should we use?

First, let's talk about **Language**. Language plays a crucial role in the process of fake news classification pipeline. Each language has its own properties: its own vocabulary, its own grammar and its own conventions. Text mining techniques, especially the ones that act upon content base features, depend on functions and methods that were built for certain types of languages. For instance, there are methods that are based on words morphology - the way that a word is built - allow to group different words that mean the same into a single common word (e.g., lemmatization and stemming techniques) in order to facilitate the process of classification. Other methods allow to detect syntactic (e.g., subject, predicate, vocative) and semantic (e.g., sentiment analysis, opinion extraction) values from a text. Hence, the success of any text mining task is highly correlated with the mature status of the previously mentioned methods and functions, in a specific language. There are, already, several Python packages that allow us to work with Portuguese language, however we do not know how mature they are and how they will behave when addressing the fake news problem with European Portuguese news.

In order to apply the aforementioned text mining techniques and to understand what kind of limitations Portuguese language imposes, we need the most fundamental asset - **Data**, Portuguese data. Here, there were two options:

- **Classify news by ourselves:** when a scientific, academic or a group of people in a company do not have the data they need, they usually create it by their own. In the context of fake news, this is a very slow process since every news needs to be extracted, analyzed and classified accordingly. Also, it is not only about time and resources, the most important thing to be considered it is peer validation. We are classifying news from social media to political matters. It is important that our analysis is recognized and certified as a process that it is independent and finely honed.
- **Use news previously classified:** this is the most desired way to make things done. Data is somewhere already classified, and all we need to do is to gather all of this information into a dataset and work from there. Since there are no datasets in the European Portuguese language what we need to do is to find a place where news are already classified. This is where ERC's work comes in.

ERC (Entidade Reguladora da Comunicação Social) is a standalone entity of the Portuguese Republic that is responsible for the supervision and regulation of the different media outlets such as press, radio, television, web content and others [12]. ERC must assure the respect for all the legal and constitutional rights and duties, making sure that fundamental values such as the freedom of speech, the press freedom, the right of being informed and the right to inform, the impartiality, the transparency and the independence from political and economic powers are not forgotten. In the end, it is the authority that assures the respect and protection of the public, in particular the youngest and sensible one. ERC always act after the exhibition of a content and not before, regardless the media outlet, meaning that this standalone entity is not a censorship instrument.

In 2019, ERC published a detailed study that aims to reflect on the dimension, scope and problems surrounding the proliferation of misinformation and false online narratives, within the European and national legal framework, limited, however, to the framework of attributions and competences entrusted to them [11]. According with this entity, the proliferation of misinformation has led to the phenomenon of fact-checking. Fact-checkers, by taking responsibility for identifying misinformation, they become, in a certain way, information producers. By taking the role of "deciding" what is "true" and "false", consequently, traditional media is also subject to scrutiny. Also, every entity that assumes the nature of the media, should immediately register in ERC. In this work, ERC refers that, in Portugal, there are two newspapers that have a "fact-check" section which is dedicated to fighting fake news: Observador and Polígrafo. Both newspapers have a certified partnership with Poynter, a well recognized international fact-checking network. And this is where this work will feed off its data. Now, knowing "what" data and "why" should we use it, some fundamental questions are left to be answered: Where can we access it? How do we get this information? And how often should we get it?

After all of those questions are answered we will find our working tool, the first, or at least one of the firsts, datasets with news from Portugal.

3.1 Fact-checkers

Fact-checking has become a prominent facet of political, economic, sports and social news coverage. This task is defined by employing a variety of methodological practices, such as treating a statement containing multiple facts as if it were a single fact and categorizing it as accurate or inaccurate. These practices share the tacit presupposition that there cannot be a genuine political debate about facts, because facts are unambiguous and not subject to interpretation. Therefore, when the black and white facts, as they appear to the fact checkers, conflict with the claims produced by politicians, the fact-checkers are able to see and detect lies [26].

In the past few years, fact-checking have been a regular task that any journalist must practice in their working daily bases. It is not something they should do, but something they must do, since it is what their journalist deontological code implies. However, with the exponential growth of social media usage, and the urge of the traditional media to apply different methodologies to be able to keep up the business and compete with those new forces, in many cases that deontological code has been put aside. Also, since there is not yet legal consequences, at least in Portugal, for the fabrication and propagation of fake news (mostly due liberty of speech being a sensitive topic as we live in a democracy) they started to appear everywhere, from social media to every traditional media outlet.

With this high competition between media sectors along side with this sense of impunity, sensationalist and made up news started to emerge. Today, not only we are living in the era of data – an era marked by the privilege of being able to access a massive amount of information (more than we ever could) – which is already difficult to process, but we also are living an era of disinformation. This is why it is imperative to have solutions to allow people to be well informed. Having this in mind, many newspapers around the world decided to adapt and build teams that are fully dedicated to fact-checking duties. Their focus, as any other private company, it is also to make profit. However, the strategy that they adopt is different. They focus on the quality and reputation of the group, and that is why even knowing that this could cost more in the near-term, making sure that that the information that they dispose is reliable, they build a sense of trust between the group and the community in the longer term. As previously mentioned, in Portugal, as it was pointed out in the ERC study, there are two newspapers that have their own department of fact-checkers: **Polígrafo** and **Observador**. Both newspapers are certified by Poynter, the owner of the International Fact-Checking Network (IFCN), a unit fully dedicated to bringing together fact-checkers worldwide. This unit was launched in September 2015 in order to support a booming crop of fact-checking initiatives by promoting best practices and ex-

changes in this field [21]. With this, both newspapers mentioned above should stand for Poynter principles – Figure 3.1.

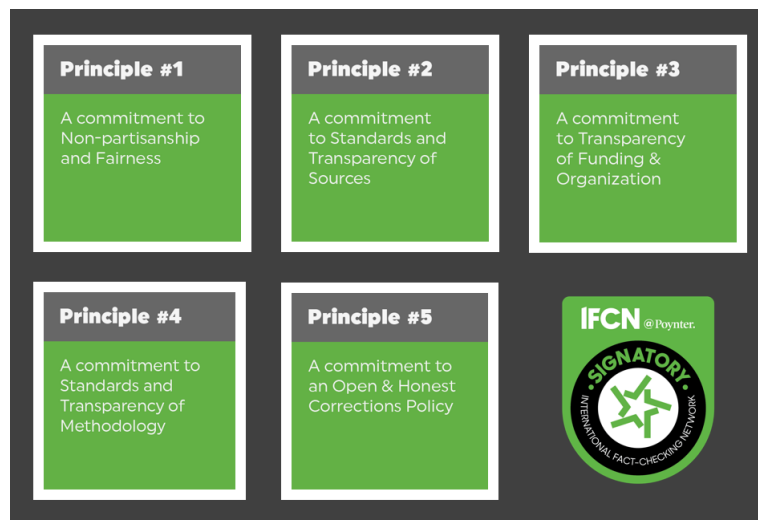


Figure 3.1: IFCN Poynter Principles

This section allow us to answer the question “where can we find our data”, mentioned in the introduction of this Chapter.

3.1.1 Polígrafo

Polígrafo¹ is a recent digital newspaper launched by Sapo². It was announced for the first time in November 2018, during the Web Summit in Lisbon. Polígrafo is the first Portuguese newspaper dealing with fake news and its database has more than two thousand classified news.

In its Editorial Status, Polígrafo presents itself as an online journalistic project whose main goal is to ascertain the truth – and not the lie – in the public space. Also, still in the Editorial Status, there are three statements that are important, if there is real commitment on them. First, Polígrafo claims a non political-ideological agenda. Journalists on the newspaper board are not militants of any political party. Second, the newspaper bases its texts on credible sources, sharing, whenever possible, links, videos, photographs, documents or other material that may contribute to clarify the ongoing discussion. And third, the newspaper does not accept anonymous sources. Between publishing an article based on an unidentified source or not publishing it, Polígrafo always chooses the second option. Those statements are adaptations from Poynter principles showed above.

The best practices of fact-checking worldwide go in the direction of, once a fact-checking task has been done, classifying its degree of veracity according to a scale. This is what ref-

¹<https://poligrafo.sapo.pt/>

²<https://www.sapo.pt/>

erence newspapers do, such as the American Politifact and Washington Post, the Argentine Chequeado or the Brazilians Agência Lupa and Aos Fatos. According to Polígrafo, reality is not white or black, so the labelling scale adopted by the newspaper has five levels (Figure 3.2):

- **Verdadeiro** (True): When the analyzed statement is totally true.
- **Verdadeiro, mas...** (True, but...): When the statement is structurally true but lacks on context and background to be fully understood.
- **Impreciso** (Imprecise): When the information contain elements that distort, even if slightly, the reality.
- **Falso** (Fake): When the statement is proven wrong.
- **Pimenta na Língua** (Tongue Pepper): The maximum level of falseness. This classification it is only applied when the information is scandalously fake or it is a satire, published in satirical space.

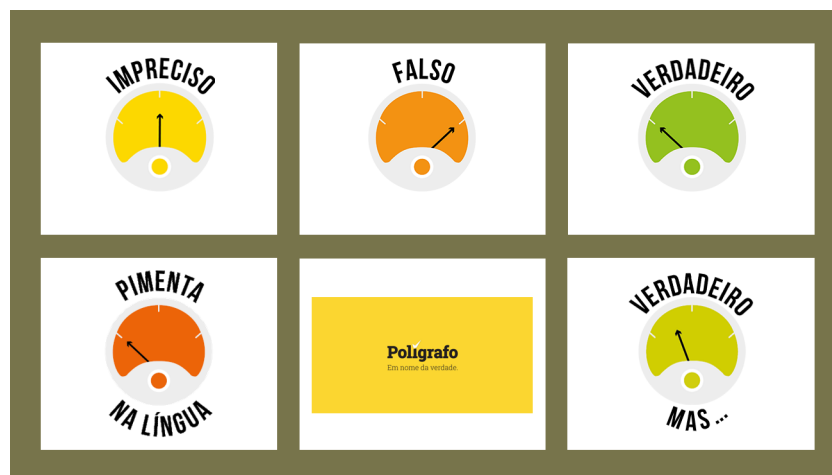


Figure 3.2: Polígrafo Tags

3.1.2 Observador

Observador³ it is also a recent online newspaper, born in 2014. It defines itself as an independent and free online daily newspaper which searches for the truth and submits to the facts. Observador stands for not being conditioned by partisan and economic interests or any group logic. They are accountable only to their readers.

Over the years, Observador have been gaining a lot of popularity and respect by the Portuguese population and even their peers. In 2015, the newspaper decided to create a section dedicated to fact-checking, and they became the first Portuguese newspaper having

³<https://observador.pt/>

a department fully dedicated to fact-checking duties. Today, Observador has more than three hundred classified news, significantly less than Polígrafo. Like Polígrafo, the fact-checking classification is not black and white. Observador has six different ways to classify news (Figure 3.3):

- **Inconclusivo** (Inconclusive): When the analyzed statement is dubious, not even a small part proven right or wrong.
- **Enganador** (Misleading): When the analyzed statement is mostly fake, and leads the reader to a misleading conclusion.
- **Errado** (Wrong): When the information is completely fake.
- **Esticado** (Uncertain): When part of the argument is true but some statements are dubious.
- **Praticamente Certo** (Almost Right): When most of the argument is true but there is a small side case the needs to be mentioned to not be misleading.
- **Certo** (Right): When the statement is proven to be right.



Figure 3.3: Observador Tags

3.1.3 Coronaverificado

During Covid-19 pandemic crisis, Polígrafo made a partnership with “Corona Verificado”, a fact-check platform coordinated by “Agencia Lupa” from Brasil, which integrates information from 34 different fact-checkers entities from 18 Ibero-American countries. This platform was launched in May 2020 and, at this moment in time, has more than two thousand news classified. Coronaverificado is the first platform of fact-checking in European Portuguese about coronavirus news and is updated in a daily basis, while the pandemic is declared by the World Health Organization (WHO).


0	Checagem Múltipla	Duvidoso	forced	Meia verdade	pepper	true-but
almost_true	Checagem Múltipla.	Enganoso	imprecise	Mentira	Precipitado	Verdadeiro, mas
Boato	Discutível	false	Impreciso	misleading	Questionável	Verificação múltipla
Boatos	Distorcido	Falso	Impreciso.	Não se sustenta	Questionável.	Viral
Certo	Dúbio	Fora de contexto	Incerto	Não verificável	Sátira	
Certo, mas	Dúbio.	Fora do contexto	Manipulado	Parcialmente falso	true	

Figure 3.4: Coronaverificado Tags

Like Polígrafo and Observador, the fact-checking classification is not black and white, and, in fact, since this platform represents an aggregation of different fact-checkers, it is also the platform in our study which has the most diversity of tags. With this said, Coronaverificado has at the moment, 40 different ways to classify news, as shown in the Figure 3.4.

3.2 Web Scraping

Now, that we know where to extract our data, the main question is: how do we get it? And this is where web scraping comes in. Web Scraping, also called “web harvesting”, “web data extraction” or even “web data mining”, can be defined as “the construction of an agent to download, parse, and organize data from the web in an automated manner” [5]. This means that, instead of having a human copying and pasting the information that he finds relevant in a web browser into, for example, a spreadsheet, web scraping handover this task to a computer program which can execute it much faster, and more correctly, than a human can. The automated gathering of data from the Internet is probably as old as the Internet itself, and the term “scraping” has been around for much longer than the web.

But, in practice, how does it work? If we want to extract information from a website, it is important to know how a website works. Briefly, every time we insert an URL into a web browser and we successfully enter in it, what we see is a combination of three technologies: HTML, CSS and Javascript. HTML “is the standard language for adding content to a website. It allows us to insert text, images, and other things to our site. In one word, HTML determines the content of a web page” Zafra [28]. And HTML alone is what we need for this part of the work.

With the help of two Python libraries made for this task, **cfscrape**⁴ to connect to the

⁴<https://pypi.org/project/cfscrape/>

website and extract the HTML and **BeautifulSoup**⁵ to parse it, we can easily code two crawlers, associated with the periodicity we want, to be able to extract the news already classified from both fact-checkers mentioned before - Polígrafo and Observador.

3.3 Database Import

Database import is the last steps of our news extraction pipeline. And why do we need it? In fact, with the previous steps mentioned in this chapter we were just fine to build our dataset, and the goal mentioned in the introduction was complete. Importing data to a database has two benefits in relation to exporting data to files, or in memory.

First, if we want to make this an asset to be re-processed and re-used every time we want, it is much better to have an unique point of access, instead of having multiple, lets say, CSV files, one for each run.

Second, web scrapping techniques have a very high potential of extracting useful information. Extracting information from a website allow us to have access to many enriched features, from the person who wrote the news to the news itself. This means that not only we have available content base features but also meta-data features, information regarding the context of the news (e.g., author, date, clicks, etc). A feature that might not sound relevant to us, might be relevant to somebody, and maybe that feature will make the difference in the classification pipeline. That is why we decided to, along with the features that we find more interesting, also import all the information regarding the HTML content, for each news. With that, everyone can use this resource from the ground basis, not being limited to what we found more relevant, allowing the community to keep up with this work.

As important as the reason behind the database import, it is the description of the database metadata. In Table 3.1, every field that belongs to this table is specified.

3.4 Dataset Overview

The next, and final step, it is the generation of the most necessary element in the Data Science domain, the dataset. This is nothing more than making use of Python to create a connection with the aforementioned database and to load in to memory all the information that we find relevant to proceed to EDA (Exploratory Data Analysis) tasks, text mining techniques and further machine learning and deep learning classification techniques that a fake news classification pipeline implies. In the end, the desired asset will be finally created – the **FakePT** dataset.

Before jumping to the next section where the classification tasks are detailed, it is important to take a look into the dataset characteristics and some EDA.

⁵<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Field ID	Field Name	Detail
1	LINK	The link to the website from which the news was extracted.
2	NEWSPAPER	News newspaper.
3	DATE	News date publication. (Not the original news, the news in the fact-checker web site)
4	CATEGORY	News category (e.g., Politics, Sports, etc)
5	TITLE	News title.
6	SUMMARY	News summary.
7	REAL_NEWS	The original news. The news that the newspaper's fact-checker classified.
8	SOURCE	The source from which the news was extracted.
9	TAG	News classification given by the respective fact-checker.
10	FACTCHECK_NEWS	News written by the newspaper analiser, regarding the original news.
11	HTML_NEWS_FACTCHECK	HTML from wich the "factcheck_news" was extracted.
12	HTML_GRID_FACTCHECK	HTML from which the all the other information besides "factcheck_news" were extracted.

Table 3.1: Database Table Fields

3.4.1 Characteristics

In terms of intrinsic characteristics there are some topics that it is important to mention. First, the dataset structure. The dataset has almost the same fields/columns/dimensions as mentioned in Table 3.1, with the exception of the following: "HTML NEWS FACTCHECK", "HTML GRID FACTCHECK", "FACTCHECK_NEWS" and "LINK". Those fields have been removed since they will only be useful for those who intend to do further investigations and decide to keep up with this work (again, bare in mind that the HTML is the source code for all the news content that have been extracted). All the other remaining fields, listed in Table 3.2 have been used in further analysis and experiments.

Dimensions		
Title	Summary	Category
Source	Newspaper	Tag

Table 3.2: FakePT Dimensions

Second, the dataset time interval. The process (i.e., from web scraping to database import) that feeds our database was designed to be easily refreshed. Every time the process runs (every time the user wants), if there are any news that were not previously imported, that new data will be imported and the database will be updated. Then, in order to proceed with a correct analysis and evaluation of our data in the following sections, we froze our extraction timeline in 2020-08-14. Our dataset is then focused in Portuguese news between June 2018 and August 2020.

Lastly, the dataset length, in terms of number of records. For this period of data, there are exactly 3764 classified news, coming from three different fact-checkers, where the respective amount of news by each one is the listed in Table 3.3.

Polígrafo	2136
Observador	594
Coronaverificado	1034

Table 3.3: News by Fact-checker

3.4.2 Exploratory Data Analysis

This section is dedicated to the exploratory data analysis (EDA) of FakePT content. EDA refers to the critical process of performing initial investigations on data in order to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights as possible from it. EDA is all about making sense of data in hand, *before getting dirty with it* [17].

There are several dimensions in FakePT that could add value to our Fake News Detection pipeline. However, not all have been considered in this study. Then, the dimensions that we found more relevant and, consequently, resulted in being selected and used in this study, are the following: news **title**, news **summary**, news **category**, news **source** and news **tag** (Figure 3.5).



Figure 3.5: News Example | Polígrafo

3.4.2.1 Temporal Analysis

Before taking a deep dive into the detailed analysis of each dimension, it is interesting to understand the investment and, consequently, the viability, which has been done by the fact-checkers. Is this work, done by each entity, constant? Has it suffered sharp decreases or increases? Lets observe and analyze Figure 3.6.

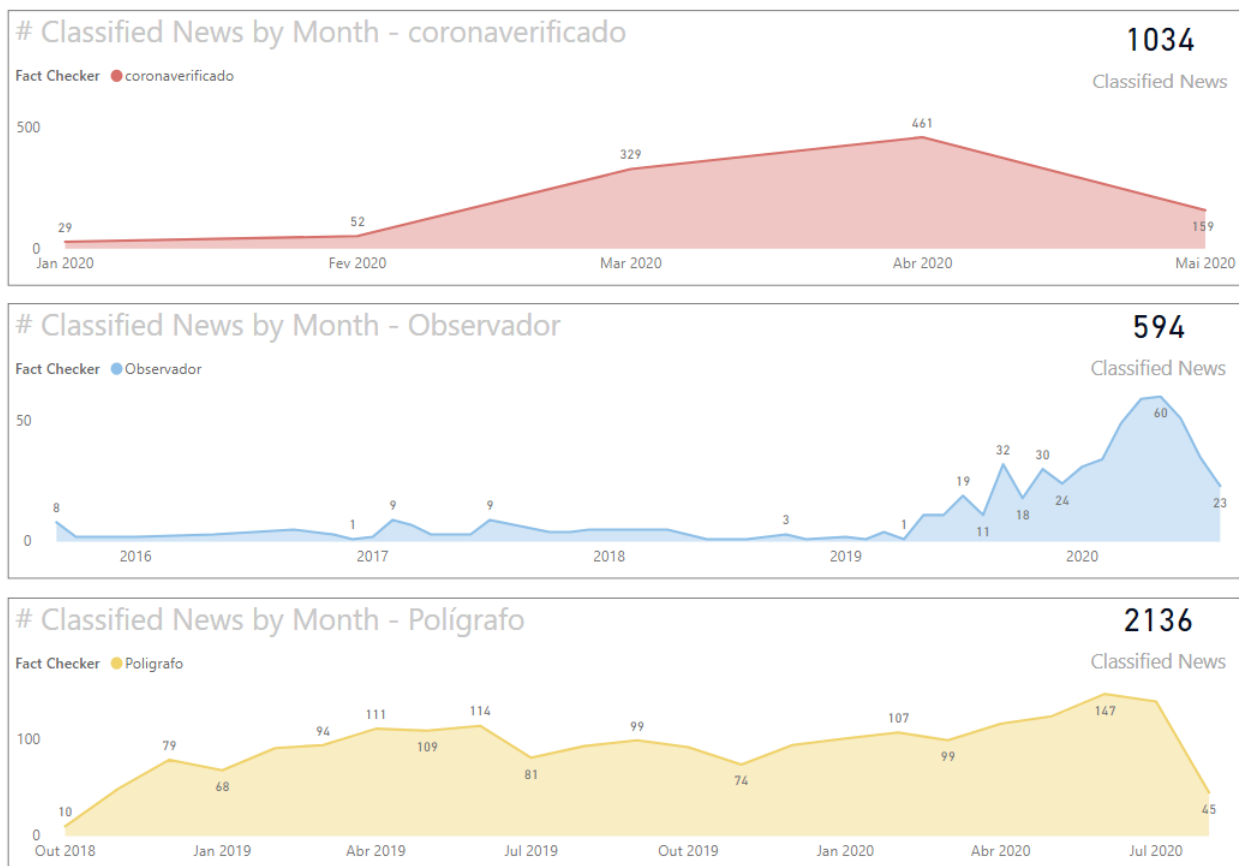


Figure 3.6: Classified News by Fact-checker and by Month

We can, through Figure 3.6 realize three different realities. The Coronaverificado presents a great contribution in relation to the total number of classified news that we have in the dataset. However, we already know that this platform was created and developed in order to exclusively guide and help the community against the avalanche of false information regarding the "Covid-19" pandemic, which began near January 2020. As such, in the longer term, this will be a source that will no longer make sense to analyze.

Second, Observador. We can see that Observador is the fact-checker which has been devoting the most time to fact-checking tasks, however, it is the one with the lowest number of classified news over time until it started to considerably grow in the second half of 2019. After that period, and especially during the pandemic, this fact-checker has been standing out as a safe and growing source, revealing itself as a valuable source to take into account

in the longer term. Also, it is important to note that in September 2020, Observador and TVI (a private Portuguese television station, which have one of the most highest number of views in Portugal) entered into a contractual partnership, which will result in a necessary and regular increase of analyzed and classified news over time.

Finally, Polígrafo. Although it has not existed for as long as Observador, Polígrafo stands out as one of the fact-checkers with the most classified news over time, since October 2018, being one of the most regular and in constant evolution. This fact-checker, like Observador, has a partnership with a private television station, but this one with SIC. This partnership has existed since 2019 and is still present today, resulting in a platform that needs to be updated, representing a reliable source in the longer term.

3.4.2.2 Dimension Analysis

This sub-section is dedicated to the exploration and discussion of different graphical representations regarding the dimensions being used in this study. In order to have an appropriate way to seize and perceive the data, different types of visualizations have been used, from tables to vertical bar charts, treemaps and wordclouds.

Title & Summary

First, title and summary. Both dimensions are represented together since they contain quite similar content, differing structurally, in most cases, in the number of words. The title, as the name indicates, is the headline for a specific news in analysis. The first thing that capture the reader's interest. Most of the time the title comes in the form of a question mark, approaching the subject of the news under analysis in a direct way: It is true or false that "something". The summary, besides the direct statement of the news theme, which is underlying the title, adds some contextual information that helps the reader to understand the context of the news. "How?", "who?", "where?", "what?" and "when?", are the typical questions that are answered in summary.

Both dimensions are textual, and, as such, unstructured. Some information that may be relevant in the future process of features selection is the number of words that makes up each news item. Thus, in Table 3.4 we can observe that the title has an average of 13.8 words per news, with a standard deviation of 5.6. On the other hand, as expected, the summary has an average of 41.1 words per news item, and an even higher standard deviation, of 18.08. The maximum and minimum word values were also calculated for each dimension.

	Title	Summary
Mean	13.8	41.1
Std	5.6	18.1
Min	1	3
Max	53	136

Table 3.4: Number of Words by Dimension

Category

In second, the category. Category enters here as the first contextual dimension associated with the news, associating it with a topic, i.e. category, in which it can be inserted. Certain topics may be more likely to be associated with fake news, and the category can be a valuable dimension to take into account in the fake news classification process. It is a categorical dimension and this is how it will be treated in the following chapters. Lets analyze each visual representation for each dimension, by each fact-checker.

In Figure 3.7, we can see on the vertical bar graph that there are no categories associated with the Polígrafo that are null. This is good news because it means that nulls do not have to be indirectly replaced by other values. In the treemap we can observe the distribution of the different news by the different categories, where the categories "Facebook" and "Politics", are the most highlighted in terms of quantity.

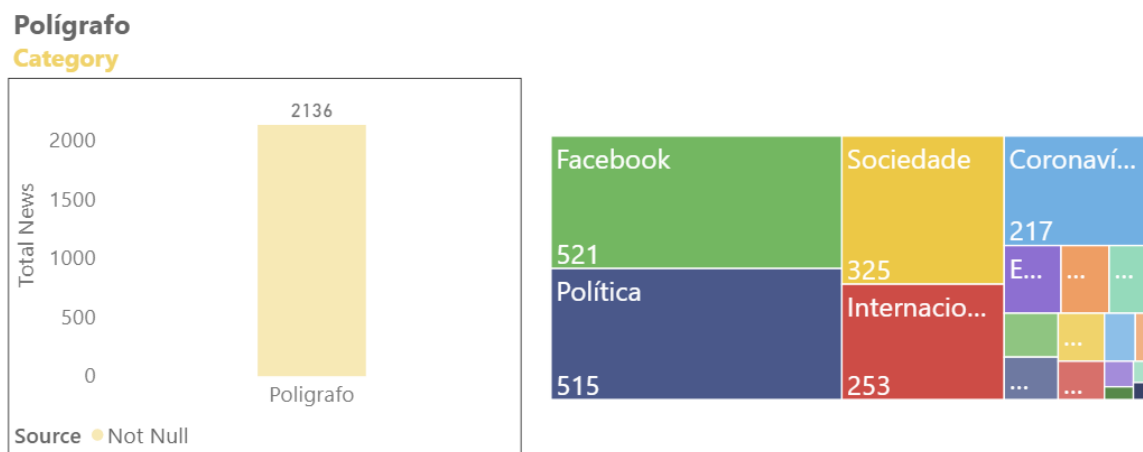


Figure 3.7: Polígrafo | Category Values Distribution

Regarding Observador, in Figure 3.8, we can see right at the start of the vertical bar graph that, this time, there are news that do not have any associated category. This happened not because there was a problem on the extracting process of the data, but because Observador does not associate a category to every news. This situation started to develop in the beginning of the Covid-19 pandemic, and that is why we can observe in the treemap that the only existing category, associated to news from Observador, is the category named "Coronavírus".

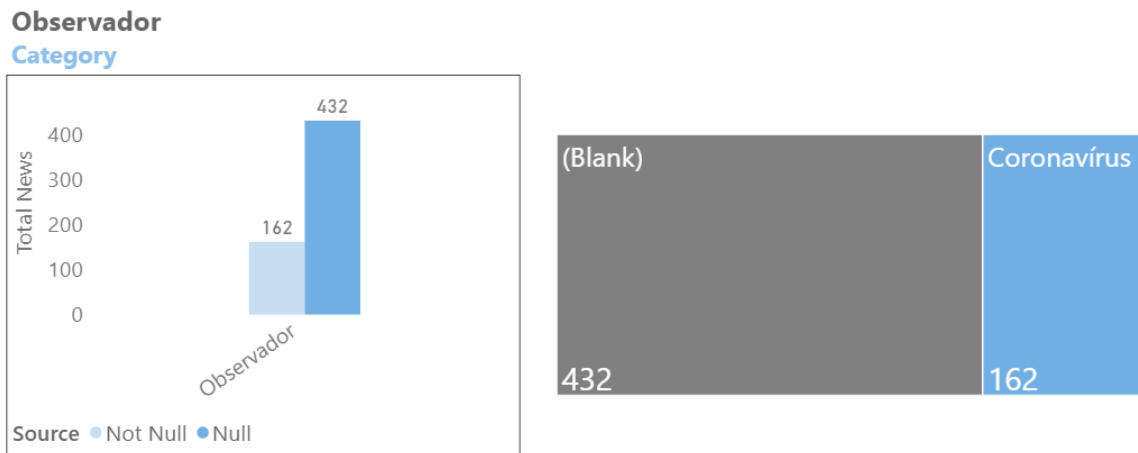


Figure 3.8: Observador | Category Values Distribution

Finally, the Coronaverificado. In Figure 3.9, we can also see that there are no category values as null. This case is particularly easy to explain as this entire platform addresses the same topic, "Coronavirus", as we can see in the treemap in the same image.

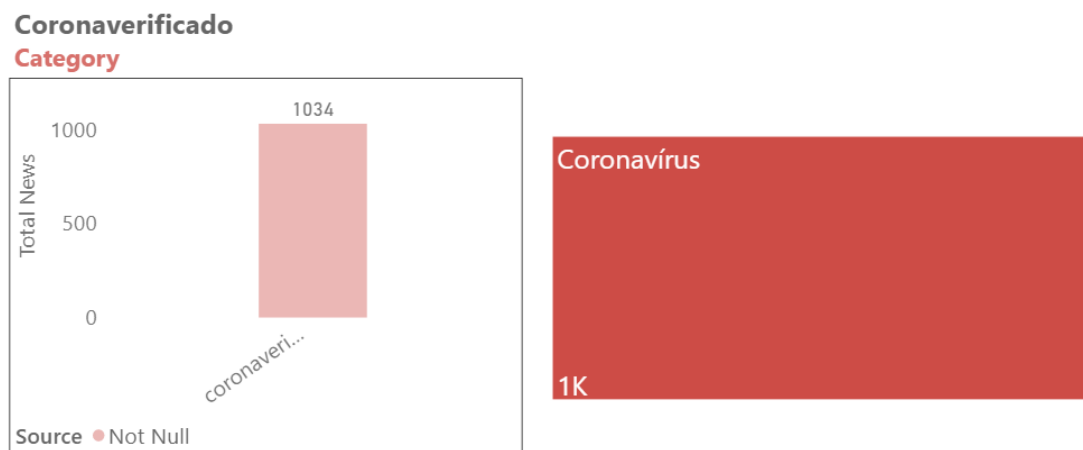


Figure 3.9: Coronaverificado | Category Values Distribution

Source

The fourth dimension is the source. Like category, it represents a contextual information of the news that can have a lot of weight in the classification of a news item. For instance, a specific less reliable source, i.e., that is associated with more false news, can be a decisive factor in the classification of a future news item not yet classified. The goal is to make this dimension categorical, however, due to the high textual noise associated with the source of each news, this dimension will be treated as textual in order to be analyzed in the next visualizations and, later, it will be worked on and converted into categorical.

Starting, once again, with Polígrafo, in the bar graph in Figure 3.10, we can see that

most of the news do not have any associated source (2048 news with source vs 88 without source). Again, such behavior is not due to the source processing, but to the absence of the association between a source and the news itself, in the fact-checker website. Therefore, for those cases it will be necessary to measure these sources indirectly. This procedure is detailed in the next chapter, in Section 4.1.1. However, we can see in the wordcloud of the same Figure the representation of the other few existing sources.



Figure 3.10: Polígrafo | Source Values Distribution

Moving on to Observador, we can see on the bar graph of Figure 3.11, that all news from the Observador have an associated source. However, as we can see in the wordcloud of the same image, these same sources do not come in the desired format, i.e., categorical. This dimension will therefore have to be processed, in order to make it as uniform and aggregate as possible. This procedure is also detailed in the next chapter, in Section 4.1.1.

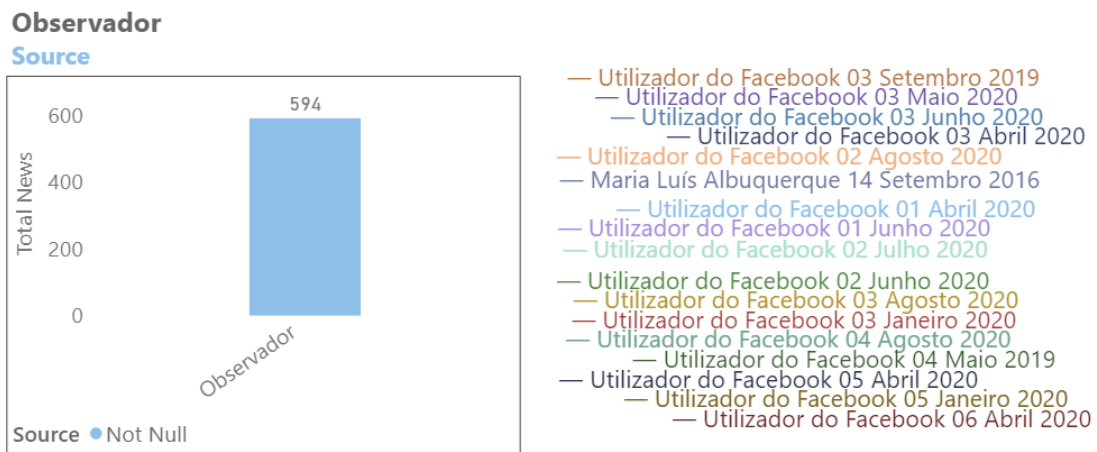


Figure 3.11: Observador | Source Values Distribution

Finally, the Coronaverificado presents most of the news (1019 news with source vs 15 news without source) with an associated source, as we can see in the bar graph of Figure 3.12. However, as we can see in the wordcloud of the same Figure, this information is very

scattered, also revealing the need of some kind of cleaning and data processing.



Figure 3.12: Coronaverificado | Source Values Distribution

Tag

The fifth and last dimension is the Tag and this represents the "Target" of the fake news classification pipeline. As previously mentioned, in the analysis of the different fact-checkers, we count with a quite large number of tags. In the dataset there are forty different tags, however, as we can see in Figure 3.13, the news density associated to each one of them is not balanced at all. There are many tags that are associated with a small number of news. This analysis, and consequent data processing, it is also explained in the next chapter, in Section 4.1.1.

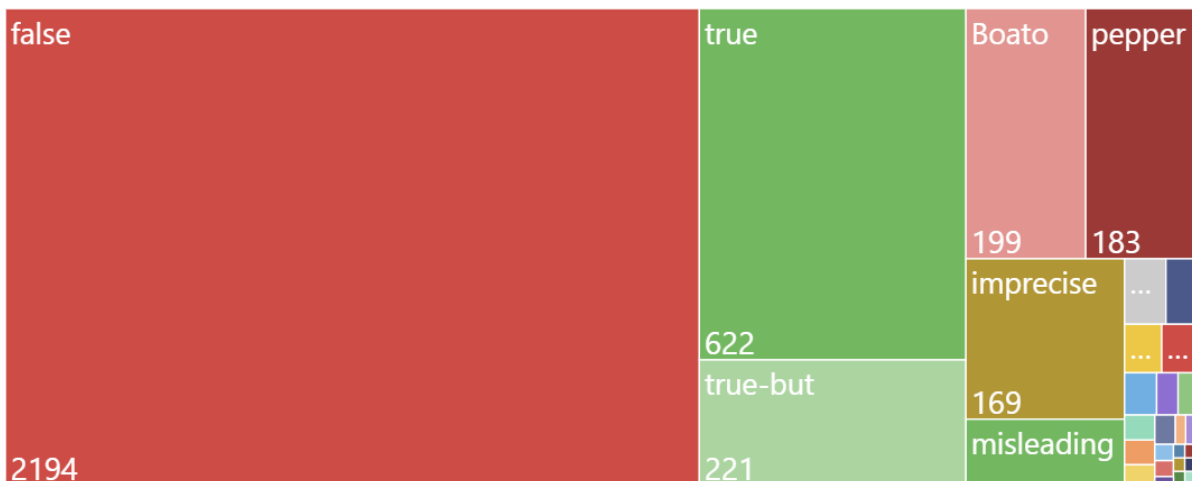


Figure 3.13: All News | Tag Values Distribution

4

Fake News Detection

In this chapter all the tasks that have been done regarding Data Pre-Processment, Data Selection and Data Classification will be detailed. In order to succeed with the implementation of most of the aforementioned approaches in the summary table of Section 2.4, all the tasks represented in Figure 4.1 have been done.

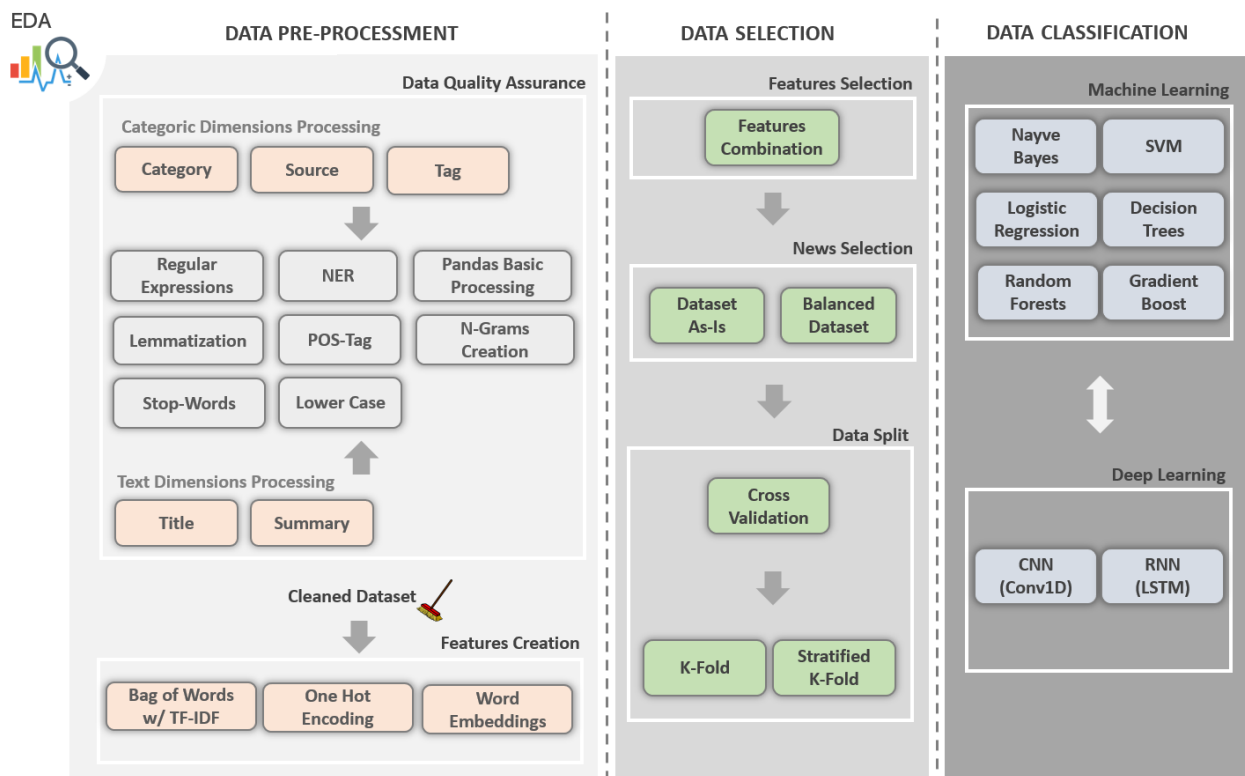


Figure 4.1: Fake News Classification Pipeline

4.1 Data Pre-Processment

This is the first module of Fake News Detection Pipeline. In here, all the necessary tasks to assure the quality of the different dimensions are described. Basic and NLP techniques were used in contextual/categorical dimensions (i.e., Category and Source), in intrinsic/text dimensions (e.g., Title and Summary) and in the target/categorical dimension, the Tag. In

the end, the strategy applied to the features creation is also explained. Those are the features that will then feed the Data Split module. This module is divided in two main groups of tasks: Data Quality Assurance (divided in “Categorical Dimensions Processing” and Text Dimensions Processing”) and Feature Extraction.

4.1.1 Data Quality Assurance

Exploratory data analysis is a crucial step in every data science challenge. Not only it provides the creation of a longer and wider picture of the data but it also allows to spot empty or poorly populated dimensions in our dataset. Removing news which happen to not have (e.g., nulls), or do have but not in a desired way, certain meta data, is not a good practice because it might result in a loss of a big chunk of data. In order to make use of the maximum potential of the dataset, some techniques of Data Quality Assurance need to be implemented. Then, FakePT needs to be cleaned first and, only afterwards, we can proceed to the creation of the different features.

In Chapter 3, an exploratory analysis was made by each dimension. All dimensions need to have specific concerns, or techniques applied, depending on the data that they represent. There are three main groups of dimensions in FakePT: structured categorical dimensions, unstructured categorical dimensions, and unstructured text dimensions. To deal with each group of dimensions, different techniques were used, from Basic (e.g., nulls replacement, data cleaning, data fill, regular expressions, tokenization, lower case, stop words removal) to NLP techniques (e.g., lemmatization, named-entity recognition, POS-Tag, n-grams creation).

4.1.1.1 Categorical Dimensions Processing

Next, all of the techniques applied are specified by each categorical (structured and unstructured) dimension. It is important to mention that all the data quality assurance process steps for each dimension, does not necessary impact the others. Many tests have been done with the combination of different features from different dimensions, which means, for instance, that if we drop one thousand null records from “Category”, if the features that are being used for a specific test are from “Source” and “Title”, if the pre-eliminated records are not empty for those dimensions, they will not be dropped. In Section 4.2 all of those experiments are explained.

Category

As it was stated in Table 4.1, there are 594 classified news from Observador, 2136 from Polígrafo and 1034 from Coronaverificado in FakePT. However, not all news have their metadata in good shape. In Table 4.1, we can see that happening for “Category” dimension.

In order to guarantee the maximum number of records as possible, in a cleaned and mature way, it is necessary to realize a set of different processes. It is important to also mention that is not always possible to fill nulls, and sometimes there is no other way but their removal. In Table 4.1 all the different processes are specified. In this dimension two different steps have been done. First, for Polígrafo only, data that might be written in a different way, but represent the same thing, have been normalized. Lastly, for Observador, since we have 432 nulls and no way to fulfill this data with other data (at least with the methods and techniques selected to apply in this dissertation), all the nulls have been dropped. One point for future work, in this topic, could be the exploration of some techniques which might help to find a category based on title and summary, like topic recognition.

	Process	Example	Not Null Records (Before)	Not Null Records (After)
Polígrafo	Normalization - Basic Pandas Data Processing	Coronavirus -> Coronavírus	2136	2136
Observador	Nulls Drop (432 records)		162	162
Coronaverificado			1034	1034

Table 4.1: Category Cleaning Steps by Fact-checker

Source

The next dimension to process is “Source”. Source, according to the literature, is one of the most important metadata regarding fake news classification. An entity, such as an organization or a person, that constantly has news classified as being fake, might create a strong pattern and a valuable input for machine and deep learning algorithms. In Table 4.2, all the processes of data quality assurance applied to FakePT are specified.

Starting by Polígrafo, first, a source assumption has been made. This means that all the news, which are associated with a “Category” like “Facebook”, now have “Facebook” also as a “Source”. Here, “Facebook” can be considered either a category or a source, since all Polígrafo news that are associated with a “Facebook” category, have their origin in that exact social network. Then for “Social Networks Detection” a simple Pandas¹ (a library in Python) data process have been applied into “Title” and “Summary” dimensions. The goal here is to have a list of conditions to find specific words that might point to a source, such as “Social Network”, “Facebook”, “Twitter”, etc. Afterwards, the first NLP technique was applied. For “Entities Detection” NER (Named Entity Recognition) has been applied. With this technique we intend to find, also in “Title” and “Summary”, words that represent an entity such as a “Person” or a “Organization”. For this task, a module of Spacy² library,

¹<https://pandas.pydata.org/>

²<https://spacy.io/>

already trained using a set of Portuguese news, has been used. Lastly, also using NER, we spotted all the sources that have a specific person or organization and classified as its entity type, i.e., “Person” or “Organization”. In this approach we decided not to have the real person and organization names because it could lead to a wide variety of total “Sources”, leading to few news with more than one common source, and, in the end, to cause a poor algorithm performance due to possible overfits. Polígrafo as it is stated in Table 4.2, before the process, started by only having 88 news associated with a source. After the process it ended up with 2130 news (2136 news in the dataset from Polígrafo) with a source.

For Observador, only two processes took place. Observador, contrary to what happened in Polígrafo, has no null values. However, all sources come in a format that needs to be normalized. That is why that for this fact-checker, only the “Social Networks Detection” process and “Entities Detection” process, using NER, have been used. Also, in the “Social Networks Detection” process, the detection was not made using “Title” or “Summary”, but “Source” itself.

Lastly, for the Coronaverificado news source, the exact three last processes that have been applied in Polígrafo, took place. In this case, despite having way more news with a source associated, compared to Polígrafo, the exact same steps need to occur. Coronaverificado as it is stated in Table 4.2, before the process, started by having 1019 news associated with a messy source. After the process it ended up with 1019 news (all news in the dataset from Coronaverificado) with a source normalized.

Tag

As it was mentioned in Chapter 3, if we take into account all the news from all the fact checkers, we have more than thirty different tags in our dataset. In order to have a truth-labeling classification (i.e. being able to find if a news it is fake or not) we need to do some data normalization. In Figure 4.2 the rational behind the said normalization is presented. In our approach we grouped and re-classified all the news tags depending on some conditions.

With that said, if a news is proven to be true, it will be classified as True. Also, if there is no proof that a news is slightly manipulated, distorted or even entirely fake, then it will also be classified as True. Lastly, if a news is structurally true, but lacks some context or background to be fully understood, we still consider it True since the integrity of the news is not compromised.

On the other hand, if a news is not structurally true due to the presence of some elements that can distort, even if slightly the reality, then it will be classified as False. Also, if there is a proof that the news is entirely fake, it will be obviously classified as False. Lastly, satires. Satires are not a consensus subject. In this study we consider only the ground veracity of the news. Even if in a theoretical assumption we should consider them neither

	Process	Example	Not Null Records (Before)	Not Null Records (After)
Polígrafo	Source Assumption	If category = 'Facebook' then source = 'Facebook'	88	2130
	Social Networks Detection	If 'Facebook' in ('Title' or 'Summary') then source = 'Facebook'		
	Source Normalization Named Entity Recognition	If 'Person' or 'Organization' in ('Source') then source = 'Person' or 'Organization'		
	Entities Detection Named Entity Recognition	If 'Person' or 'Organization' in ('Title' or 'Summary') then source = 'Person' or 'Organization'		
Observador	Social Networks Detection	If 'Facebook' in ('source') then source = 'Facebook'	594	594
	Entities Detection Named Entity Recognition	If 'Person' or 'Organization' in ('source') then source = 'Person' or 'Organization'		
Coronaverificado	Social Networks Detection	If 'Facebook' in ('Title' or 'Summary') then source = 'Facebook'	1019	1034
	Source Normalization Named Entity Recognition	If 'Person' or 'Organization' in ('Source') then source = 'Person' or 'Organization'		
	Entities Detection Named Entity Recognition	If 'Person' or 'Organization' in ('Title' or 'Summary') then source = 'Person' or 'Organization'		

Table 4.2: Source Cleaning Steps by Fact-checker

true nor false, but what they really are, satires, in our practical context we should consider them fake, since in the end they cannot be considered as true.

After this process we have 872 (23%) true news in our dataset, and 2889 (77%) fake news. Those numbers were obtained before any pre-process steps over the different dimensions, except Tag.

4.1.1.2 Text Dimensions Processing

This section is dedicated to the data quality assurance regarding unstructured text dimensions. Contrary to what happens to categorical dimensions, where the main goal is to have one news associated with a simple word, or a short set of words, resulting in a short vocab-

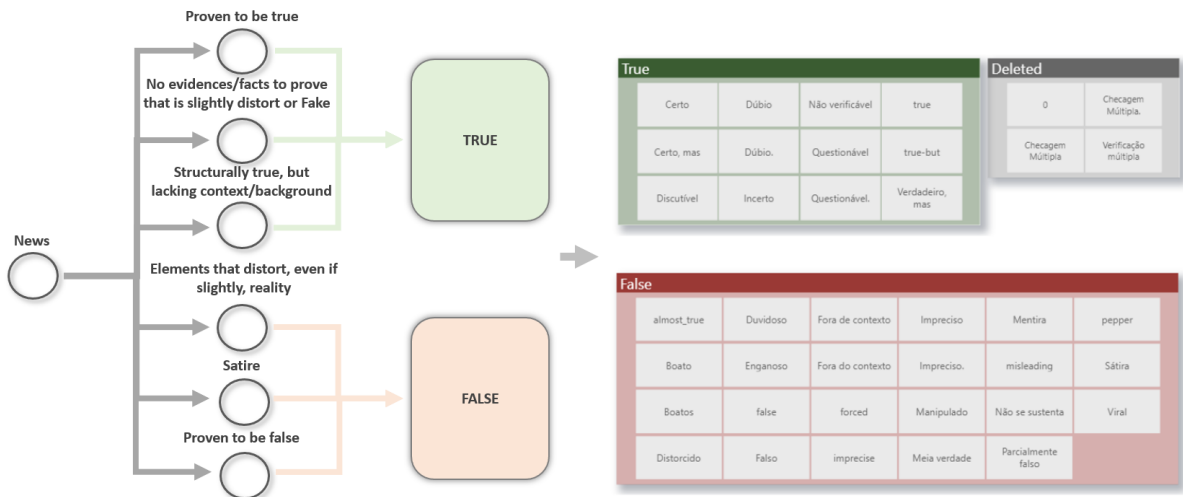


Figure 4.2: Tags Classification Assessment

ulary of different categories, the idea behind unstructured text dimensions is quit different. The main goal for those dimensions is to capture some kind of meaning from each text dimension. By exploiting in-depth news text analysis, we can analyze linguistic patterns and writing styles for both truth news and fake news, and then capture the most discriminative features for online fake news detection. The current studies on news text analysis can be categorized as: linguistic and semantic-based analysis, knowledge-based analysis, and style-based analysis [29].

- Title & Summary: the main dimensions in FakePT. Many tests will be done in order to understand which are the most meaningful techniques that allow us to have the best outcome when applying different algorithms, in terms of evaluation. However, there are some tasks that cross every test:
 1. The use of regular expressions to eliminate punctuation;
 2. Tokenization;
 3. The elimination of a specific vocabulary of stop words (e.g., words with one character, words or expressions that often appear [VÍDEO, Verdade ou mentira?, Será verdade?, etc]);
 4. Creation of n-grams;
 5. Lowercase all the words.

Despite the above techniques, as it was stated before, there are some that are only used on occasional test situations such as Named-Entity Recognition, POS-Tag and Lemmatization. All the experiments and respective results are presented in the next Chapter.

4.1.2 Feature Extraction

This module ends up with the features creation. This is the last step before data selection, and one of the most important ones in the Fake News classification pipeline. The main goal here is to use the cleaned dimensions mentioned in the previous section and create the necessary features from them, using different dimension representations techniques. This brings up two challenges:

1. Which type of representation should we use for each dimension?
2. Is it possible to combine them? If so, how?

Starting by challenge one, like it was mentioned in the previous section, we have different type of dimensions in our dataset: text and categorical ones. For text dimensions, like it was mentioned in Section 4.1.1, we want to be able to read and transform a lot of unstructured data into a set of relevant features. One token (one word, or a combination of n-words) will result in one feature. Then, each news title or summary will be described as a set of features. For this case we have decided to use two different text representation approaches: **Bag of Words w/ TF-IDF**, and **Word Embeddings** (both described in previous literature). As for categorical dimensions, the type of representation needs to be different. For their representation we used a technique named **One Hot Encodding**, wich means that every news categorical dimension will be represented as a sparse vector, filled with zeros and one number one. Each vector will have as many values as the number of categories which describe each dimension, and each news will have only one number “one”, which is associated with the category associated with that news and all the other values filled as zero.

In challenge two, we combined different categorical and textual features in different ways, depending on the text representation approach we want to use. For the Bag of Words w/ TF-IDF approach a method from Sklearn³ library named “ColumnTransformer” was used. For the Word Embeddings approach, we created different input layers, one for each of the desired features to use, and combined them.

4.2 Data Selection

This is the second module of Fake News Classification Pipeline. Here we explain the rational behind our feature selection approach, what data did we use and how did we split it into train and test.

³<https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html>

4.2.1 Feature Selection

The idea behind this task is to be able to test as many valuable combinations of features as possible, in order to find the best input to the classification task. In Table 4.3, all the tests that have been done, for the different set of features and different features representations are described.

Test	Set of Features	Features Representation
1	Title	BoW w/ TF-IDF
2	Summary	BoW w/ TF-IDF
3	Title + Summary	BoW w/ TF-IDF
4	Best Approach from Test (1,2,3)	Word Embeddings
5	Best Approach from Test (1,2,3) + Category	BoW w/ TF-IDF + One Hot Encoding
6	Best Approach from Test (1,2,3) + Source	BoW w/ TF-IDF + One Hot Encoding
7	Best Approach from Test (1,2,3) + Category	Word Embeddings
8	Best Approach from Test (1,2,3) + Source	Word Embeddings

Table 4.3: Feature Selection

4.2.2 News Selection

Also, as it was stated in Section 4.1.1, we do not have a balanced dataset. With this in mind, we decided to duplicate all the previous mentioned tests in order to be able to compare two different approaches:

- Tests using a balanced dataset (50% fake news vs 50% true news).
- Tests using an unbalanced dataset (77% fake news vs 23% true news).

For the second approach it was selected, randomly, as much fake news as there is true news on FakePT dataset. Then, a new dataset was created with the same number of fake and true news.

4.2.3 Data Split

The last step of this module is data split and this is an essential preparatory part before jumping to the classification task. Here data is split in train and test. The most common approach it is to use a simple method for data splitting, where the whole cleaned dataset is divided in two, only once, with values between 70% and 80% for the training set and values between 30% and 20% for the testing set. Machine Learning models often fail to generalize well on data they have not been trained on. That is why we need a more complex method

for data splitting. To be sure that the model can perform well on unseen data, we use a re-sampling technique, called **Cross-Validation**. In our study we used two similar, but still different Cross-Validation techniques: **K-Fold** and **Stratified K-Fold**.

K-Fold gives a model with less bias compared to other methods. In K-Fold, we have a parameter 'k'. This parameter decides in how many folds the dataset is going to be divided. Every fold gets the chance to appear in the training set (k-1) times, which in turn ensures that every observation in the dataset appears in the testing set, enabling the model to learn the underlying data distribution better. The value of 'k' used is generally between 5 or 10. The value of 'k' should not be too low or too high. In our case, we used a 'k' value of 5. This approach is useful for balanced datasets, so we are going to use it in the first approach mentioned in Section 4.2.2.

Another approach is to shuffle the dataset just once prior to splitting the dataset into k folds and, then, split such that the ratio of the observations in each class remains the same in each fold. Also the test set does not overlap between consecutive iterations. This approach is called Stratified K-Fold. This approach is useful for unbalanced datasets, so we are going to use it in the second approach mentioned in Section 4.2.2.

4.3 Data Classification

This is the last module of Fake News Classification pipeline – data classification. In Section 2.3, we highlighted several classifiers that have been used in the literature, within the scope of Fake News. In this section we present all the algorithms that have been selected, for all the experiments. Both machine learning and deep learning classifiers were used – Table 4.4.

Machine Learning				Deep Learning
Multinomial Naive Bayes (MNB)	Random Forest (RF)	Logistic Regression (LR)	Extra Trees (ET)	ConV1D (CNN)
K-Nearest Neighbors (KNN)	Linear SVM (LSVM)	SVM (SVM)	Gradient Boosting (GB)	LSTM (RNN)

Table 4.4: Classifiers

5

Results

Here we discuss the results obtained for the different set of experiments for the FakePT News Classification Pipeline. We evaluate the effectiveness of the ten classifiers specified in Section 4.3 for news classification by cross-validation on the FakePT dataset, and then choose the best performing one based on different data conditions.

In the end, all these data conditions and respective best classifier method (the one with better score results for one or more evaluation metrics), are compared, in order to understand what are the dimensions, the data pre-processing, the text representation techniques, and the classifiers that lead to better evaluation scores.

5.1 Evaluation Metrics

Many evaluation criterias are used for assessing the performance of different machine and deep learning techniques. The most common metrics are the following: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision (Pr), Recall (Re), F-score (F1) and Accuracy (Acc). The first four ones, TP, TN, FP, FN, are detected fake news, detected true news, misclassified true news, and undetected fake news, respectively. Also, the formulas for calculating the last evaluation metrics are the following:

$$Pr = \frac{TP}{TP + FP}$$

$$Re = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Depending on the data that the algorithms are being applied, different types of evaluation metrics makes sense to use. Typically, in binary classification scenarios, like finding if a news is true or false, Accuracy is mostly used when the dataset is balanced. In unbalanced scenarios, other evaluation metrics like F1-score, Precision and Recall are the most fit to use. More detailed explanations for the aforementioned metrics can be found in [Jiawei2012]

5.2 Experiments

This section is dedicated to detail the experiments. In this data classification type of problems, there are way too many variables that are interesting to use and susceptible to change. Variables like dimensions selection, types of text processing techniques, types of categorical processing techniques, types of text representation, different set of news selected and the classifiers selection and their hyper-parameter tuning. With so many variables, some decisions need to be made regarding their selection approach and, inevitably, some experiments will be left out.

With this in mind, the experiments took place in two different phases: one phase to find the best set of text processing techniques (basic and NLP) and, the last one, to find the best combination of criterias (i.e., set of features and best classifier) for each type of news selection approach – Section 4.2.2.

5.2.1 Phase 1

Phase 1 of the experiments aims to assess which is the best combination of textual processing techniques, using three different sets of features, and one specific classifier. The combinations of variables that could be made by joining the experiments of both phases would be enormous, and so this step serve as the basis assumption for the following phase. Thus, the best set of techniques identified in this phase will be the one that will be used in all the experiments of phase 2.

Lets begin by describing Table 5.1. The first column of the table refers to the set of textual processing techniques applied in the different experiments. Individually, the techniques that are identified in this column are the following:

- Punctuation removal;
- Lemmatization;
- Stopwords removal;
- Lowercase;

- N-grams spacy (1): the ability to detect if a word, or a set of words, represent an entity. Here, the “pt_core_news_sm”, a pretrained statistical model from spacy, which assigns context-specific token vectors, POS tags, dependency parses and named entities, is used. The possible entities that are spotted here are: “MISC”, “ORG”, “PER”, “LOC”.
- N-grams spacy (2): same as (1) but, here, the possible spotted entities are: “ORG”, “PER”, “LOC”.
- N-grams TF-IDF (3): the goal is the same as the two previous points, but this time it is “TfidfVectorizer”, a model from sklearn, which is responsible for the creation of the n-grams. N-grams and the maximum number of features (words or set of words) used to represent all the news also are variables that vary from one to three, and from one thousand to ten thousand, respectively.

The second column refers to the different textual dimensions that were used in the experiments: Title, Summary and Title plus Summary. It is important to emphasize that here, as well as the technique of creating n-grams through TF-IDF, different maximum values of number of features were always tested for different algorithms, from one thousand to ten thousand. Finally, in the third column, the average accuracy of the different iterations of the k-fold is presented. The accuracy is used here, to the detriment of other evaluation measures, since the set of news, being used in this classification, is balanced in terms of target class distribution – **balanced dataset** (50-50).

Also, for all the experiments described in Table 5.1, it was always used the same text representation, **BoW w/ TF-IDF**, and the same machine learning classifier – **Multinomial Naïve Bayes**.

5.2.2 Phase 2

It is in the second and last phase of experiments where all the final evaluations for our Fake News Classification Pipeline are presented. All the work developed in this dissertation comes down to this final stage, where all of the final evaluation metrics are shown for the different combinations of variables presented in Table 5.2.

Moving on to the detailed description of the contents of Table 5.2. In the first place, the “Dataset” column. Here the possible values in the different tests are “Balanced” and “Unbalanced”. With this separation we intend to conclude which are the best evaluation metrics, and the conditions that gave them origin, of course, for each of these datasets.

In second, the dimensions. However, here there are two differences in relation to the dimensions of the previous table. For each of the datasets (balanced and unbalanced) three different tests are performed, and in each one the only thing that changes is the dimension used. These three tests have the objective of checking which set of features (generated through the dimensions Title, Summary or Title plus Summary, using BoW w/

Text Processing Techniques	Dimensions	AVG Acc
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (1); Lowercase	Title	0.63
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (2); Lowercase		0.64
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams TF-IDF (3); Lowercase		0.65
Punctuation Removal; Lemmatisation; POS-Tag; Stopwords removal; N-grams TF-IDF (3); Lowercase		0.64
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (1); Lowercase	Summary	0.68
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (2); Lowercase		0.69
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams TF-IDF (3); Lowercase		0.71
Punctuation Removal; Lemmatisation; POS-Tag; Stopwords removal; N-grams TF-IDF (3); Lowercase		0.69
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (1); Lowercase	Title + Summary	0.68
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams spacy (2); Lowercase		0.69
Punctuation Removal; Lemmatisation; Stopwords removal; N-grams TF-IDF (3); Lowercase		0.69
Punctuation Removal; Lemmatisation; POS-Tag; Stopwords removal; N-Grams TF-IDF (3); Lowercase		0.70

Table 5.1: Experiments | Phase 1

TF-IDF) presents the classifier with the best evaluation metrics, within the various machine learning classifiers under study. Thus, the dimension that presents the best results will be the one that will be used in later experiments, together with the categorical dimensions.

In third place we have the different textual representations. The different textual representations are the following:

- Bag of Words w/ TF-IDF;
- One Hot Encoding;
- Word2vec FakePT: a skip-gram 100 dimensions model created based on the vocabulary of FakePT text dimensions;
- Word2vec NILC: a skip-gram 100 dimensions model developed by the “Núcleo Interinstitucional de Linguística Computacional”, which was trained over different data sources [9].

Next, we have the column referring to the best classifier, "Best Classifier", the one within the 10 classifiers used in test (2 for deep learning and 8 for machine learning). It is important to mention that deep learning algorithms were only used in conjunction with the features created based on word embeddings. Finally, the column referring to the average score of the different metrics, Accuracy (Acc), Precision (Pr), Recall (Re) and F1-score (F1).

5.3 Discussion

The last section of this chapter is dedicated to the discussion of the different experiments and respective evaluation outputs. As it was mentioned in Section 5.2.1, the goal of Phase 1 was to create a text processing baseline in order to facilitate and reduce the number of upcoming experiments, by reducing the number possible variable combinations. Therefore, after finding the best test processing techniques, Phase 2 starts as described in Section 5.2.2, where the main goal is to compare different classifiers performance, for different input variables.

Phase 1

A clear difference can be seen between dimensions, regarding the performance of the MNB classifier, despite the type of text processing techniques used in Table 5.1. For all of the same text processing experiments by dimension, “Title” had always a worst Average Accuracy performance, comparing to “Summary” and to “Title + Summary”. The best set of text processing techniques for Title is “Punctuation Removal; Lemmatization; Stopwords removal; N-grams TF-IDF [3]; Lowercase”, which led to an average accuracy of **0.65**, which

Dataset	Dimensions	Text Representation	Best Classifier	Average K-fold Score						
				Acc	False			True		
					Pr	Re	F1	Pr	F1	
Balanced	Title	BoW w/ TF-IDF	SVM	0.66	0.66	0.63	0.64	0.64	0.66	
	Summary		ET	0.74	0.78	0.67	0.72	0.71	0.76	
	Title + Summary		SVM	0.70	0.70	0.69	0.69	0.69	0.70	
	↓ EVALUATION (Best Dimension) ↓									
	Summary + Source	BoW w/ TF-IDF	ET	0.74	0.78	0.69	0.73	0.72	0.76	
	Summary + Category	One Hot	ET	0.71	0.70	0.75	0.72	0.73	0.70	
	Summary	Encoding Word2vec (FakePT)	RNN	0.67	0.68	0.64	0.66	0.66	0.68	
	Summary + Source		CNN	0.69	0.70	0.67	0.68	0.68	0.69	
	Summary + Category		RNN	0.68	0.67	0.64	0.66	0.65	0.68	
	Summary	Word2vec (NILC)	CNN	0.64	0.65	0.63	0.64	0.64	0.65	
	Summary + Source		RNN	0.66	0.67	0.64	0.68	0.66	0.67	
	Summary + Category		RNN	0.64	0.65	0.62	0.64	0.63	0.64	
Unbalanced	Title	BoW w/ TF-IDF	LSVM	0.79	0.82	0.92	0.87	0.57	0.43	
	Summary		LSVM	0.78	0.83	0.90	0.86	0.53	0.45	
	Title + Summary		LSVM	0.80	0.83	0.93	0.87	0.60	0.45	
	↓ EVALUATION (Best Dimension) ↓									
	Title + Summary + Source	BoW w/ TF-IDF	LSVM	0.79	0.84	0.89	0.86	0.58	0.52	
	Title + Summary + Category	One Hot	MNB	0.80	0.87	0.85	0.86	0.68	0.60	
	Title + Summary	Encoding Word2vec (FakePT)	CNN	0.79	0.84	0.89	0.86	0.55	0.50	
	Title + Summary + Source		CNN	0.80	0.88	0.87	0.86	0.54	0.51	
	Title + Summary + Category		CNN	0.78	0.83	0.85	0.84	0.52	0.48	
	Title + Summary	Word2vec (NILC)	RNN	0.79	0.84	0.89	0.86	0.55	0.48	
	Title + Summary + Source		RNN	0.79	0.83	0.88	0.87	0.53	0.50	
	Title + Summary + Category		RNN	0.79	0.82	0.86	0.86	0.51	0.48	

Table 5.2: Experiments | Phase 2

is worst than the worst performance of the same techniques, when applied to the other dimensions (0.68 average accuracy for both dimensions). This bad performance or, at least, worst performance when compared to the other text dimensions, might be justified by the average length of Title. As it was mentioned in the Sub-Section 3.4.2.2, Title has an average number of words of 13.8 when Summary has an average number of words of 41.1. Summary has almost four times more words than Title, which means that the classifier has way more information to work with, in order to understand the meaning behind each text.

On the other hand, Summary and “Title + Summary” are way more similar in terms of evaluation outputs. Summary, despite having the best score (**0.71** average accuracy), does not stand out from “Title and Summary”, with **0.70** of average accuracy.

Having said this, the selected set of text processing techniques for the Phase 2 are “Punctuation Removal; Lemmatization; Stopwords removal; N-grams TF-IDF [3]; Lower-case”, which led to an accuracy of **0.71**, using the **Summary** dimension, **Bag of Words with TF-IDF** as the approach for text representation, a balanced dataset, and the **Multi-nomial Naive Bayes** (MNB) as the classifier.

Phase 2

In the next phase, there are two main set of experiments: the ones made with a **balanced** dataset, and the ones made with an **unbalanced** dataset. The goal is the same for both of them – to find the best combination of features and classifiers which better adapts and, consequently, better perform under the data in hand. However, the way of evaluating the goal’s performance is going to be different in each case. Also, like it was explained in Section 5.2, there are many combinations that can be done in each step of the Fake News Classification Pipeline. Then, a “Dimension” selection is also done in each of the two main set experiments, which means that the dimension (or dimensions) which lead to the creation of the features that better perform, are the ones that are selected for further experiments. In this phase, only the three text dimensions are compared: “Title”, “Summary” and “Title + Summary”.

In the case of the balanced dataset, the metric that is going to be used in order to compare performances, is **Accuracy**. Like it was mentioned in Section 5.1, accuracy is the best metric to evaluate the performance of an algorithm when a balanced dataset is used. Looking at Table 5.2, the best accuracy score out of the three experiments with different text dimensions is **0.74**. This means that 74% of the predictions were right, using “**Summary**” as the main dimension, **Bag of Words w/ TF-IDF** as the text representation technique, and **Extra Trees** as the classifier. Summary was selected has the main text dimension to further experiments and it was combined with different categorical dimensions, under different text representation techniques. The experiments were done and the best score, using **Bag of Words w/ TF-IDF**, goes to the combination of “**Summary**” and “**Source**”, using the **Extra Trees** classifier, which also led to a score of **0.74**. As for the different word

embeddings techniques, the results were not that promising, since the best score obtained was **0.69**, when combining **Summary** and **Source**, **word2vec (FakePT)**, and **CNN**.

On the other hand, for the unbalanced dataset, the metric that is going to be used in order to compare performances is **F1-score**. Since it is easy for a classifier to overfit under a very uneven tag distribution, recall, precision and f1-score are the metrics that we should look up to. F1-score is the metric that seeks to find a balance between Precision and Recall, and the one that should be used when there is an uneven class distribution [25]. Moving to the experiments, we can see in Table 5.2 that **“Summary + Title”** is the text dimension with a better performance of **0.87** F1 score in the **“False”** class and **0.45** in **“True”** class, using **Bag of Words w/ TF-IDF**, **One Hot Encoding** and **LSVC**. Due to this result, **“Summary + Title”** is the selection dimension for the following experiments. After this, only two best scores are relevant to be highlighted. The first one, and the one that outperformed the other classifiers using an unbalanced dataset, goes to the combination of **“Title + Summary”** and **“Category”** as the main dimensions, **Bag of Words w/ TF-IDF**, **One Hot Encoding** and **Multinomial Naive Bayes**. This combination resulted in a **0.86** F1 score in the **“False”** class and **0.60** in **“True”** class. Lastly, the best performing algorithm using word embeddings (**word2vec - FakePT**), was achieved by using a combination of **“Title + Summary”** and **“Source”** dimensions and by using a **CNN**. This combination resulted in a **0.86** F1 score in the **“False”** class and **0.51** in **“True”** class.

6

Conclusions and Future Work

6.1 Conclusions

This work presents one of the first studies in the context of Fake News classification in European Portuguese and, also, an alternative approach in the way that news are typically treated and classified in this kind of problems. The main objective of this work is to present a valid, interesting and trustworthy way of classifying Portuguese news, which can be used as another tool, or in a complementing way to other approaches, to help anyone who wants to fight this new phenomenon of misinformation called Faked News.

This was the first work, at least according to the survey we made, which addresses the classification of news in European Portuguese. In order for this to happen, through a web scraping mechanism, 3764 news from different sites dedicated to the task of fact-checking were collected, and then stored in a relational SQL database, i.e., sites that have specialized teams dedicated exclusively to the analysis of the truthfulness of the arguments that make up each of the most controversial news found in traditional and social media, focusing in the social, sports, political and economic weekly agenda. These news are already classified in this way, and each site adopts its own classification and labeling rationale regarding the veracity of each one of them, but always following the norms and guidelines of Poynter, the owner of the International Fact-Checking Network (IFCN).

However, the classification of these news implies an added concern since this is not a typical Fake News classification problem. In all the approaches we identified in the literature, the features that were used in the classification tasks were diverse, from the features generated through the news content itself (content-based features) to the features generated from the context (context-based or source based features). In our study, although features generated from the context can be comparable with other works, features generated from the content itself cannot be directly compared, since the news we have for analysis are news created over other news. Thus, here we do not try to apply text mining and classification techniques to the source news, but rather to news written by fact-checkers.

From the different dimensions that characterized each of the news, the dimensions "Title", "Summary", "Category", "Source" and "Tag" were selected as the features we proposed to analyze, which resulted in the construction of our dataset – FakePT. As far as

pre-processing is concerned, different techniques were applied depending on the type of variable under analysis – textual or categorical. Regarding the processing of textual variables it was important to understand the state of maturity of the different NLP techniques existing in the Portuguese language. Here, surprisingly, there were already enough libraries to handle different parts of the pre-processing, from the removal of Portuguese stop words (models from NLTK¹ library), to the use of models that allow the use of tokenization, lemmatization and POS-Tagging (models from spacy and Stanza² libraries) and the creation of word embeddings using models already created on Portuguese content (word2vec NILC). However, many of these models had been trained on content in Brazilian Portuguese and it was still unclear how they performed on news in European Portuguese.

Many were the experiments done on the aforementioned dimensions, from the types of textual representations (word embeddings and bag of words w/ TF-IDF) and categorical (one hot encoding) used, to the different types of classification techniques (machine learning and deep learning). Also, to observe the behavior of each of these algorithms, tests were made for a balanced dataset and for an unbalanced dataset.

The results presented, referring once again to not being able to be directly compared with the works in literature, were quite promising. In the balanced dataset the best accuracy score was **0.74**, using "**Summary**" as the main dimension, **Bag of Words w/ TF-IDF** as the text representation technique, and **Extra Trees** as the classifier. In the case of the unbalanced dataset, "**Title + Summary**" and "**Category**" were used as the main dimensions, **Bag of Words w/ TF-IDF** and **One Hot Encoding** as text and categorical representations respectively, and **Multinomial Naive Bayes** as the classifier. This combination resulted in a **0.86** F1 score in the "False" class and **0.60** in "True" class.

Lastly, this methodology has proved to be quite interesting for future explorations, whether in academic or professional environments, being a potentially ambitious tool to help in the fight against Fake News.

6.2 Contributions

Adding to this dissertation, we also published and presented an article in the "9th Symposium on Languages, Applications and Technologies (SLATE 2020)".

The article, named "**Towards the Identification of Fake News in Portuguese**" [23], was made in parallel with this work, and presents the current state-of-the-art of this dissertation topic and some suggestions towards the future of fake news classification in Portuguese.

The article is available in "Dagstuhl Research Online Publication Server", in the following link: https://drops.dagstuhl.de/opus/frontdoor.php?source_opus=13020

¹http://www.nltk.org/howto/portuguese_en.html

²https://stanfordnlp.github.io/stanza/available_models.html

6.3 Future Work

By implementing our end-to-end solution for Portuguese European news extraction, and applying all of the experiments within our Fake News Classification pipeline, we found out some promising assets, either in research and application perspectives. However, it is clear that improvement is still needed concerning many aspects of the discussed problems.

There were a number of limitations and also a considerable number of other experiments that could have been done in our study, in order to improve the results outcome.

Data sources present the main considerable limitations. Despite being a good way of getting previously classified news in Portuguese, some news **lack some context information**. Many do not have an associated source, and many others do not have a category. In our work, we presented several approaches to infer the information that is missing, but we did not test other techniques that also might be interesting, such as topic detection, to be able to be more accurate in our prediction. Also, it is important to note that **this work have been done not with the real news**, but with the news from the fact-checker. An interesting approach would be to try to add the real news to the system, opening a whole new world of text processing possibilities to infer some of the news writer behavior. Also, the **tag assessment** concern. In most of the cases, the number of news classified as true compared to the number of news classified as false is not balanced, while most of the fact-checkers tend to analyze controversial statements, which, in most of the times, tend to be false. This, if not well treated, might result in model overfit.

From **text pre-processing**, to **text representation**, ending up in **classifiers**, there is a lot that can be done. There are always other text processing techniques and other combinations of techniques which might result in a better text representation. Also, playing with the categorical dimensions, such as tag. Transforming a binary-class problem into a multi-class problem, or, just adopting different types of news labeling assumption. Lastly, other word embedding models could have been used, in pair with other classification methods. Still, even with the used models in this approach, a lot of tuning could have been performed in order to get better performing evaluations.

Bibliography

- [1] Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. Ed. by New Riders. 2016.
- [2] Monther Aldwairi and Ali Alwahedi. “Detecting fake news in social media networks”. In: *Procedia Computer Science* 141 (2018), pp. 215–222. ISSN: 18770509. DOI: [10.1016/j.procs.2018.10.171](https://doi.org/10.1016/j.procs.2018.10.171). URL: <https://doi.org/10.1016/j.procs.2018.10.171>.
- [3] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–236. ISSN: 08953309. DOI: [10.1257/jep.31.2.211](https://doi.org/10.1257/jep.31.2.211).
- [4] Renan Rocha De Andrade. “Utilização de técnicas de aprendizado de máquina supervisionado para detecção de Fake News”. In: (2019).
- [5] Bart Baesens and Seppe Broucke. *Web Scraping for Data Science with Python*. CreateSpace Independent Publishing Platform, 2017, p. 256. ISBN: 1979343780.
- [6] Sean Baird, Doug Sibley, and Yuxi Pan. *Talos Targets Disinformation with Fake News Challenge Victory*. 2017. URL: <http://blog.talosintelligence.com/2017/06/talos-fake-news-challenge>.
- [7] Alessandro Bondielli and Francesco Marcelloni. “A survey on fake news and rumour detection techniques”. In: *Information Sciences* 497 (2019), pp. 38–55. ISSN: 00200255. DOI: [10.1016/j.ins.2019.05.035](https://doi.org/10.1016/j.ins.2019.05.035).
- [8] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. “News in an online world: The need for an “automatic crap detector””. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. ISSN: 23739231. DOI: [10.1002/pra2.2015.145052010081](https://doi.org/10.1002/pra2.2015.145052010081).
- [9] Núcleo Interinstitucional de Linguística Computacional. *NILC - Embeddings*. 2017. URL: <http://nilc.icmc.usp.br/embeddings>.
- [10] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1.Mlm (2019), pp. 4171–4186. arXiv: [1810.04805](https://arxiv.org/abs/1810.04805).
- [11] Entidade Reguladora para a Comunicação Social. *A Desinformação - Contexto Europeu e Nacional*. 2019, p. 78. ISBN: 8745970726.

- [12] ERC. *ERC Web Page*. 2020. URL: <https://www.erc.pt/pt/sobre-a-erc>.
- [13] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by Gibbs sampling". In: *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference 1995* (2005), pp. 363–370. DOI: [10.3115/1219840.1219885](https://doi.org/10.3115/1219840.1219885). URL: <https://nlp.stanford.edu/software/CRF-NER.html>.
- [14] Neema Kotonya and Francesca Toni. "Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection". In: *Proceedings of the 6th Workshop on Argument Mining* (2019), pp. 156–166. DOI: [10.18653/v1/w19-4518](https://doi.org/10.18653/v1/w19-4518).
- [15] Shuaipeng Liu, Shuo Liu, and Lei Ren. "Trust or Suspect? An Empirical Ensemble Framework for Fake News Classification". In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia* (2019), pp. 1–4. URL: <http://www.wsdm-conference.org/2019/wsdm-cup-2019.php>.
- [16] Rafael A. Monteiro et al. "Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11122 LNAI (2018), pp. 324–334. ISSN: 16113349. DOI: [10.1007/978-3-319-99722-3_33](https://doi.org/10.1007/978-3-319-99722-3_33).
- [17] Prasad Patil. "What is Exploratory Data Analysis?" In: *Medium* (2018). URL: <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15%7B%5C%7D0A>.
- [18] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), p. 4.
- [19] Lam Pham. "Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News". In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia* (2019). DOI: [10.1145/nnnnnnn.nnnnnnn](https://doi.org/10.1145/nnnnnnn.nnnnnnn). URL: <https://doi.org/10.1145/nnnnnnn.nnnnnnn>.
- [20] Dean Pomerleau and Delip Rao. *The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news*. 2017. URL: <http://www.fakenewschallenge.org/>.
- [21] Poynter. *Poynter Web Page*. 2020. URL: <https://www.poynter.org/>.
- [22] Julio Reis and André Correia. "Supervised Learning for Fake News Detection". In: (2019), pp. 76–81. DOI: [10.1109/MIS.2019.2899143](https://doi.org/10.1109/MIS.2019.2899143).
- [23] João Rodrigues, Ricardo Ribeiro, and Fernando Batista. "Towards the identification of fake news in Portuguese". In: *OpenAccess Series in Informatics* 83.7 (2020), pp. 1–7. ISSN: 21906807. DOI: [10.4230/OASICS.SLATE.2020.7](https://doi.org/10.4230/OASICS.SLATE.2020.7).

- [24] Giovanni C Santia and Jake Ryland Williams. "BuzzFace : A News Veracity Dataset with Facebook User Commentary and Egos". In: *Twelfth International AAAI Conference on Web and Social Media* (2018), pp. 531–540.
- [25] Koo Shung. "Accuracy, Precision, Recall or F1?" In: *Medium* (2018). URL: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>.
- [26] Joseph E. Uscinski and Ryden W. Butler. "The Epistemology of Fact Checking". In: *Critical Review* 25.2 (2013), pp. 162–180. ISSN: 08913811. DOI: [10.1080/08913811.2013.843872](https://doi.org/10.1080/08913811.2013.843872). URL: <http://dx.doi.org/10.1080/08913811.2013.843872>.
- [27] Ke Wu, Song Yang, and Kenny Q Zhu. "False Rumors Detection on Sina Weibo by Propagation Structures". In: *IEEE 31st international conference on data engineering* (2015).
- [28] Miguel Zafra. "Web Scraping news articles in Python". In: *Medium* (2019). URL: <https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558>.
- [29] Xichen Zhang and Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion". In: *Information Processing and Management* 57.2 (2020), p. 102025. ISSN: 03064573. DOI: [10.1016/j.ipm.2019.03.004](https://doi.org/10.1016/j.ipm.2019.03.004). URL: <https://doi.org/10.1016/j.ipm.2019.03.004>.