UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CRISTIANO RUSCHEL MARQUES DIAS

Towards fake news detection in Portuguese: New dataset and a claim-based approach for automated detection

Work presented in partial fulfillment
of the requirements for the degree of
Bachelor in Computer Science

Advisor: Prof. Dr. Dante Augusto Couto Barone

Porto Alegre
July 2019

ABSTRACT

The spread of non-veridic information is a longstanding problem that has affected society ever since the advent of communication. The emergence of the Internet and Social Media have aggravated this issue, leading to a higher degree of influence of misinformation in peoples opinions and lives and therefore a higher impact on contemporary events, such as the 2016 US Elections. This led to the coining of the term Fake News, which has been widely used describe this recent phenomenon, and prompted it's study by many fields of knowledge, and in the context of many languages. In the field of Computer Science, the main concern is the outstanding problem of automated fake news detection, which has barely been explored in the context of lusophone countries. One of the reasons for this, is the lack of content - datasets - which are required for work to be done on the subject. This work aims to provide a new labeled dataset for the problem of fake news detection in Portuguese, with news claims gathered from unbiased and non-partisan fact-checking sources, and apply methods of text-classification which have been proved to work on fake news, in order to validate if news can be classified as fake based solely on their claim. Apart from existing methods, this work also attempts a novel classification method, using the named entities extracted from the claim as a feature for classification.

Keywords: Fake news. data mining. text classification. machine learning. natural language processing.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

SVM	Support Vector Machine
LSVM	Lagrangian Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
EXIF	Exchangeable image file format
API	Application Programming Interface
NLTK	Natural Language Learning Toolkit

CONTENTS

1 INTRODUCTION

1.1 Defining fake news

The subject of this work, that of fake news, is one that is hard to define. Though many attempts have been made, no formal definition to the term is currently widely accepted by the academic community. CORNER et al describes fake news as ä kind of a fraudulent media productïn which ẗhe negative judgement and the sense of intention are even stronger than with, say, 'bias' or even with Chomsky's distorting, propagandistic 'filters'.̈ On the other hand, GELFERT goes to great lengths to create a widely acceptable definition for the phenomena, stating that "Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design".

Due to the nature of this work, a less strict definition to the term will be provided. We will be considering fake news any content which is shared on media platforms, including digital and non-digital as well as traditional and new media outlets, which contains false information, wether it be intentional or not. This definition differs from the previously presented ones in that it does not constrain fake news to something deliberate, but rather also includes heavily biased content and unintentionally false news.

1.2 Fake News throughout history

Fake news, in the form of misinformation and yellow journalism, is not a new phenomenon. Rather, it is one that has existed since antiquity. One famous example of this is the Donation of Constantine. As the power of the Catholic Church grew during the Middle Ages, conflicts arose between the Church and the European ruling class over control of the states. At that time, the Church forged the existence a document in which the 4th century Roman Emperor Constantine the Great donated most of the Empire's western lands to the Papacy, in the figure of Pope Sylvester I (RUSSELL, 2004). This document was then in the 11th century cited by Pope Leo IX, who believing it to be true, in a letter to Michael I Cerularius, Patriarch of Constantinople. From that point onward, it was cited in many times throughout history (MIGNE, 1891).

Another more recent example are the German Corpse Factories, or Kadaververwertungsanstalt. During World War I, there was an anti-German propaganda effort in the United Kingdom to make people believe that the Germans had factories to create fat from

human beings – a good which they had a short supply of due to blockades. This heavily influenced public opinion against the German Empire, allegedly even affecting the opinion of the Chinese President Feng Guozhang (FUSSELL, 2000).

1.3 The emergence of fake news

Even though it has been around for a long time, fake news is a topic that has recently been receiving an increasingly large amount of attention both by the general public and the academical community. This begs the question: Why this interest came to be? And this is a question that can be answered in multiple ways.

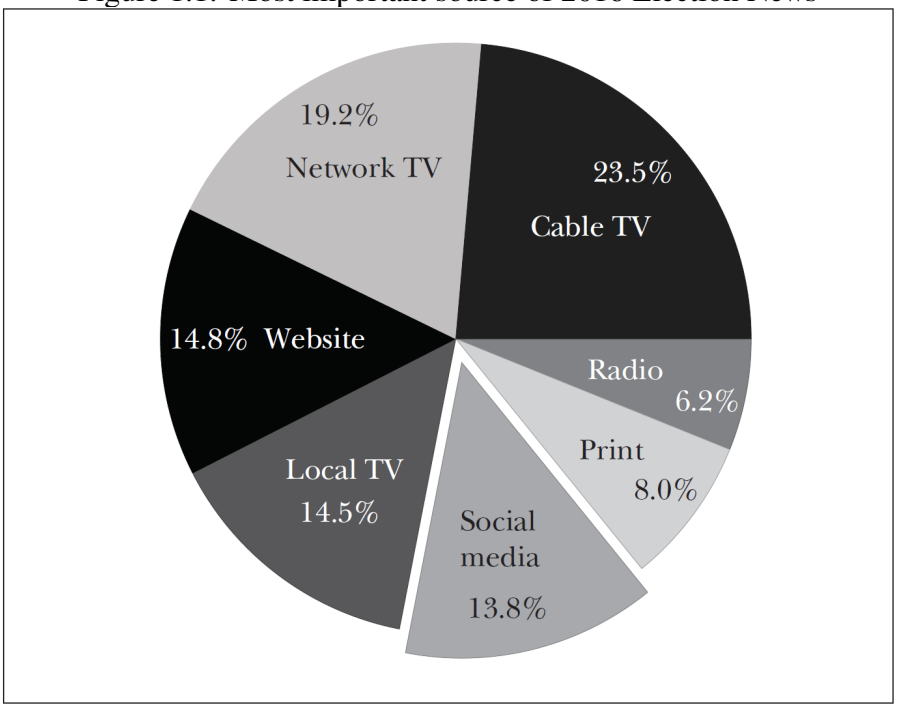
Some specialists argue that we have reached an era of post-truth politics. This is a term that was coined in 1992 by Steve Tesich on his essay in the Nation magazine (according to Oxford Dictionaries) on the following quote: "we, as a free people, have freely decided that we want to live in some post-truth world". This term defines circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. This term has been deemed the word of the year by the Oxford Dictionary in the year of 2016 [1], given it's importance in modern politics. By it's definition, it is clear to see that this post-truth reality brings forth a demand for news which are not necessarily true, but rather appeal to the masses as much as possible - fake news(POMERANTSEV, 2016).

This post-truth argument is reinforced by recent events, such as the 2016 US Presidential Elections, which have been heavily influenced by fake news. Numerous works have attempted to quantify the impact of this phenomenon on the Elections, including the survey by Allcott and Gentzkow (2017). This in specific work aimed to offer theoretical and empirical background to frame the debate on the influence of fake news on the elections. In Figure 1.1 we can see that 13.8% of the sample population responded that they mainly informed themselves about the elections through social media - one of the main environments in which fake news are generated and propagate(SHU et al., 2017). Emphasizing this, Figure 1.2 show us objectively that 8% of the respondants of the questionnaire have confirmed to have seen and believed in fake news articles related to the Elections.

Further insights on this impact are provided by GUESS A., who further show us

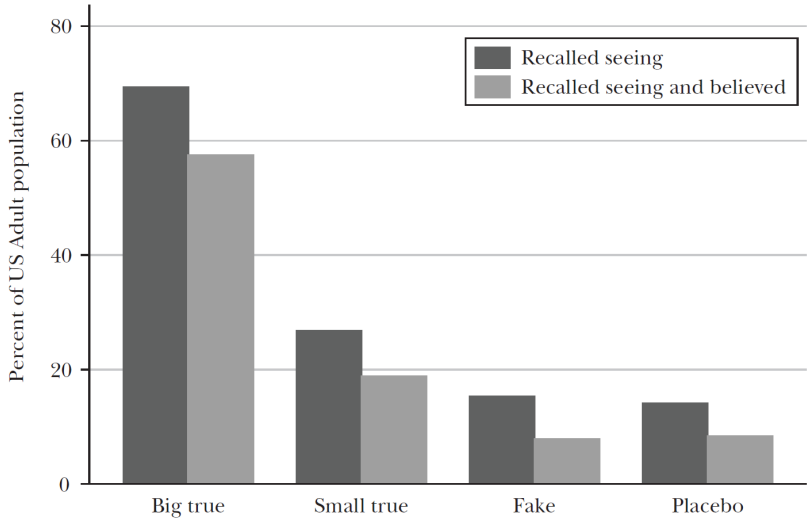
[1] Available on: https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016. Acessed on: 12 jun. 2018.

Figure 1.1: Most important source of 2016 Election News



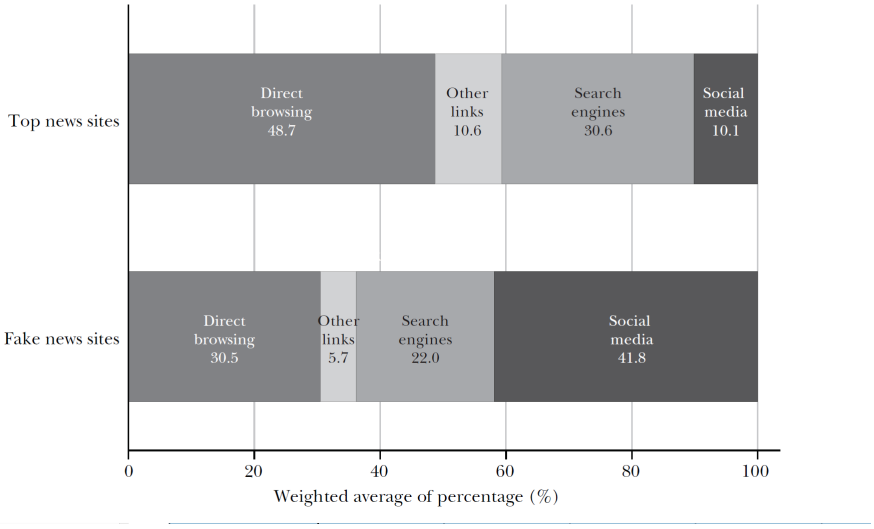
Source: (ALLCOTT; GENTZKOW, 2017)

Figure 1.2: Percent of Adult Population that Recall Seeing or that Believed Election News



Source: (ALLCOTT; GENTZKOW, 2017)

Figure 1.3: Share of Visits to US News Websites by Source



Source: (ALLCOTT; GENTZKOW, 2017)

the impact of fake news on the 2016 Elections - according to it, 27.4% of all electors (over 65 million people) visited fake news websites during the final weeks of the 2016 election campaign period. This work also brings evidence that those news were heavily targeted towards the pro-Trump audience. People saw an average (mean) of 5.45 articles from fake news websites during the study period of October 7–November 14, 2016. Nearly all of these were pro-Trump (average of 5.00 pro-Trump articles).

It is also necessary to bring to light that the advent of technologies such as the internet and social media have been key enablers of the propagation of fake news. Phenomena such as echo chambers - a situation in which beliefs are amplified or reinforced by communication and repetition inside a closed system - and filter bubbles - a medium where users see content and posts that agree only with their preexisting belief - create a perfect environment for the propagation of fake news (DIFRANZO; GLORIA-GARCIA, 2017). This is proven on a research by Pew Research, which reveals that 61% of millennials use Facebook as their primary source for political news.

The relevance of social media in the context of fake news is further emphasized by a study reported in Nature, which described a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during the 2010 U.S. congressional elections. It found the messages directly influenced political selfexpression, information seeking, and real-world voting behavior. This is all revelant as Figure 1.3 helps us see that social is the preferred propagation media of fake news.

The aforementioned factors and many others have contributed to the current interest on the subject of fake news. It is a phenomenon that requires a multidisciplinary effort to

be tackled, and as such has recently been explored in many fields such as Anthropology, Sociology and, of course, Computer Science (LAZER et al., 2018).

Specifically within Computer Science a couple of aspects and challenges regarding fake news are of particular interest. In particular, the field has given particular attention to the study of the propagation dynamics of fake news (PAPANASTASIOU, 2018) and the task of Automatic fake news detection(BOND R.M.; FOWLER, 2018).

1.3.1 This work

In this work, the problem of Fake News detection in portuguese will be explored. The first goal is to present a definition of the problem and an overview on existing works and approaches, both in other languages - predominantly english - and in portuguese, to contextualize the core of the work to be done.

In order to explore the task of automated fake news detection in portuguese, this work will provide on of the first labeled datasets of fake news in portuguese, and the first dataset to provide news claims classified by verified nonpartisan and unbiased sources.

Using this dataset, experiments will be done on claim-based fake news classification using some approaches that have been validated by literature in other contexts, adapting them and validating them in the context of fake news detection in portuguese - as well as a, to the best of my knowledge, novel approach using named entities extracted from the text as features. This part of the work aims to provide a baseline for this task and a direction for future work, and identify which features and algorithms help us achieve the best results.

2 A REVIEW ON FAKE NEWS DETECTION

On the previous chapter we defined the motivation for the topic of this work. On the following sessions, we will formally define the task of fake news detection, the approaches that have been used to attempt to solve the problem, and the state of the art on each of those approaches.

2.1 Fake News Detection

We will be borrowing an adapted version of the definition of fake news detection from Shu et al. (2017). This definition is appropriate as it enables us to make a clear separation and definition of the different approaches that are used to solve the problem.

- Let a refer to a News Article. It consists of two major components: Publisher and Content. Publisher $\vec{p_a}$ includes a set of profile features to describe the original author, such as name, domain, age, among other attributes. Content $\vec{c_a}$ consists of a set of attributes that represent the news article and includes headline, text, image, etc.
- We also define News Engagements as a set of tuples $\mathcal{E} = \{e_{it}\}$ to represent the process of how news spread over time among n sources (including social media users) $S = \{s_1, s_2, ..., s_n\}$ and their corresponding post or article $P = \{p_1, p_2, ..., p_n\}$ regarding news article a. Each engagement $e_{it} = \{s_i, p_i, t\}$ represents that a source s_i spreads news article a using p_i at time t . Note that we set $t = Null$ if the article a does not have any engagement yet and thus s_i represents the original publisher and pi the original article.

Definition (fake news detection) 1. *Given the social news engagements E among n users for news article a, the task of fake news detection is to predict whether the news article a is a fake news piece or not, i.e.,* $\mathcal{F} : \mathcal{E} \Rightarrow \{0, 1\}$ *such that*

$$\mathcal{F}(e) = \begin{cases} 1, & \textit{if a is a piece of fake news} \\ 0, & \textit{otherwise} \end{cases}$$

where $\mathcal{F}$ *is the prediction function we want to learn.*

Given the definition of the problem, we will proceed to present existing attempts at

solving the problem. These solutions can be separated in many different classes, depending on which features of the news article or engagement they use classify the news.

2.2 Machine Learning

From Definition 1, it is taken that the fake news detection problem can be seen as a binary classification problem. Solutions to this type of problem itself tend to be hard to define mathematically or be directly derived, though (SHU; WANG; LIU, 2017) offers an example of an attempt to do so. Solutions to this problem tend to make use of Machine Learning algorithms.

Machine learning is an approach which attempts to solve the problem without using explicit instructions to do so . These types of algorithms can be classified on four main types(AYODELE, 2010):

- Supervised Learning: is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- Unsupervised Learning: this machine learning task finds previously unknown patterns in data set without pre-existing labels.
- Reinforcement learning: the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- Semi-supervised learning: these algorithms are trained on a combination of labeled and unlabeled data. This is helpful as, for instance, it enables us to use larger input datasets without necessarily labeling all data.

The fake news detection, as a classification problem, is better suited to be solved by a supervised algorithm, as we are required to learn (or to approximate the behavior of) a function which maps a vector - in our definition the engagement set - into one of several classes by looking at several input-output examples of the function (KHAN et al., 2010). As such, the labeled dataset is a pre requisite to solving the problem.

An alternative to supervised learning would be to use a semi-supervised learning algorithm to solve the problem. A dataset would still be needed, but a subset of it would be unlabeled.However, current scientific research on the subject hasn't explored this approach enough to make it a topic of meaningful discussion in the context of fake news detection.

A direct consequence of this is that works on the subject need a dataset with

which to train their solution. At the moment, there are many consolidated and publically available datasets for the research of fake news in english. For instance, the FakeNewsNet dataset (SHU et al., 2017), maintained by a research group from Arizona State University, has been extensively used on many works on the subject such as (SHU et al., 2017) and (SHU; WANG; LIU, 2017) . Another commonly used dataset is the one provided by (SILVERMAN et al., 2016), which contains about 50 of the most shared fake news for each year, with their associated engagements.

2.3 Existing solutions

2.3.1 Content based

We have defined a News article as having a set of contents $\vec{c_a}$. Some of the features that may be represented in this content set are listed below:

- Source: Author or publisher of the news article
- Headline: Short title text that aims to catch the attention of readers and describes the main topic of the article
- Body Text: Main text that elaborates the details of the news story; there is usually a major claim that is specifically highlighted and that shapes the angle of the publisher
- Image/Video: Part of the body content of a news article that provides visual cues to frame the story

What we define as content-based approaches utilize only the information in the content set to attempt to classify the article as fake or not. This approach is further subdivided based on the type of content which is being used in the classification task, textual content or visual content.

2.3.1.1 Linguistic approach

These approaches utilize the textual content of the original article as the only information of interest when attempting to generate the target function. This is possible as fake news articles generally have textual cues in them that allow us to clearly distinguish them from veridic texts, such as inflamatory language or "clickbait" (CHEN; CONROY; RUBIN, 2015). Though the writer normally avoids it, some of those cues may "leak" into

the final text, especially certain textual aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage (FENG; HIRST, 2013a). The goal in this approach becomes then to look for instances of leakage, or so called “predictive deception cues” (CONROY; RUBIN; CHEN, 2015).

Given this definition, the problem becomes a classical text classification problem, which can solved by a multitude of algorithms, each exploring different features of the textual content. For instance, by using a TF-IDF representation of the corpus with stop-word removal and stemming data preparation steps and using the LSVM (MANGASARIAN; MUSICANT, 2001) classification algorithm (an implementation of SVM), Ahmed, Traore and Saad (2017) have achieved an accuracy of 92% on the fake news detection problem.

A variation of this strategy uses not the original content of the news, but only the claims made by the news themselves. The work by (KONSTANTINOVSKIY et al., 2018) makes a comparison of a series of algorithms in the context of automated fact checking, and achieves an 83% F-score while classifying claims.

It is necessary to note that these solutions use a bag-of-words representation of the text: as they represent only the frequency of each word, it does not take into account syntax nor semantics or other properties of the text.

There are many works that have achieved great result in deception detection by using other textual features, such as (NEWMAN et al., 2003) which used LIWC (Pennebaker et al., 2007), a dictionary which categorizes words or word-stems in categories such as Psychological processes, Relativity and Standard linguistic dimensions, to add semantic and syntatic information into the classification model.

By using even more textual information in the context of computerized deception detection, Feng, Banerjee and Choi (2012) have achieved an accuracy of 91.2%. They combined the simple word representation of the other approaches with what they call Deep syntax: "encodings of production rules based on the Probabilistic Context Free Grammar (PCFG) parse trees". This approach is improved even further by Feng and Hirst (2013b).

2.3.1.2 Visual approach

Nowadays it is really easy to fill scenes (HAYS; EFROS, 2007), generate composite fake images (TSAI et al., 2017), and even generate video based on speech (SUWAJANAKORN; SEITZ; KEMELMACHER-SHLIZERMAN, 2017). Due to this, image content has come to play a big part on the spread and apparent credibility of fake news [1].

[1] https://www.wired.com/2016/12/photos-fuel-spread-fake-news/

This can be seen on the work by Gupta et al. (2013), in which from a sample of 16117 tweets about the Hurricane Sandy that hit the US in 2012 which contained image URLs, some 10,350 tweets contained fake images. Just as the texts, often contain inflammatory or emotionally appealing content.

The work by Jin et al. (2016) further emphasizes the importance of visual cues in detecting fake articles, and provides an algorithm that classifies the images as fake or not by searching for visual cues.

There are also approaches that do not directly look for the visual deception cues, but rather use metadata from the images or textual description of the images to put their credibility in check. (HUH et al., 2018) presents an approach that uses EXIF metadata available on image files, which contains information such as camera manufacturer, model, configuration settings at the time of capture, among others, to look for potential image splices, a type of image manipulation in which you insert a part of another image into the original one. Using this approached, they achieved an accuracy of over 90% in recognizing tampered images.

2.3.2 Propagation based approach

In our definition of the problem of fake news detection, we specified the input of the classification function to not only contain the contents of the news, but rather the set of all engagements with said piece of news. This approach attempts to use the information contained in the propagation of the news amongst users and sources to detect deceptive cues, often comparing the propagation dynamics of fake news to that of reputable news.

VOSOUGHI; ROY; ARAL Show us that there is indeed a difference in propagation dynamic between fake and truthful rumours. In this study, false claims reached far more people than the truth - While truth rarely propagated to more than 1000 people, the top 1% of false-news cascades often reached between 1000 and 100,000 people (Fig 3.3B). Not only false claims travel further, but they also spread faster. As it can be seen on (Fig. 2F), truth claims took about six times as much as falsehood to reach 1500 people, and 20 times as long to reach a propagation depth of 10 (Fig. 2E).

(Kwon et al., 2013) Further emphasizes this findings in a different study. It shows us that indeed, fake news travel much more quickly and farther than truthful news. It provides an attempt at solving the fake news detection problem, by creating a "friendship network as the induced subgraph of the original follower-followee graph induced by those

users who posted at least one related tweets and follow links among them", and using this as one of the inputs for the classification function.

Another example of such a work is (LIU; WU, 2018), which provides a solution that uses CNN combined with a specific type of RNN, Gated Recurrent Units (GRU) (Chung et al. 2014) to achieve over 90% accuracy using datasets with over 6900 news from Twitter and Weibo.

2.3.3 Hybrid approach

Finally, we have solutions which do not fall in any single of the previously mentioned categories. Those hybrid approaches use a combination of the features earlier described and others to achieve higher levels of accuracy in the fake news detection problem.

One such example is the aforementioned work by Kwon et al.. Besides using the follower-followee graph, they also use a collection of over 80 features, between rumor diffusion patterns over time (propagation features), the shape of the diffusion network and the friendship network (also propagation features), and the language used in the content (linguistic features) to achieve their highest accuracy of 90%.

(SHU; WANG; LIU, 2017) Provides us another example of classification using multiple features. In this work, three types of features are taken into account: A representation of the bias of the publisher, the actual news content, and the sequence of social engagements for a given piece of news. They model the classification problem as an optimization problem, and prove that the different features used contribute differently and positively to the solution of the problem in their solution.

It is worth noting that, on many cases, simpler approaches lead to a result as good as approaches that take into account more features - this leads to the conclusion that more features does not necessarily lead to a better classification result. This can be seen when comparing Feng, Banerjee and Choi (2012) and Kwon et al.. Whereas the former uses only linguistic features, it achieves a higher accuracy than Kwon et al., with it's multitude of features.

3 FAKE NEWS AND AUTOMATED DETECTION IN PORTUGUESE

3.1 The prominence of bots and fake news in Brazil

While fake news as a topic of research - especially within Computer Science - is a new one, there is significant literature already available on the topic for English based texts and datasets. Meanwhile, the study of fake news in the context of Portuguese speaking countries is in its infancy.

Just like the recent events of the 2016 US Elections and United Kingdom's 2016 Brexit served as a wake-up call to the impact of fake news in shaping public opinion, the same happened recently in Brazil with the 2016 Impeachment process and the 2018 General Elections. Surprisingly though, the impact of internet robots and fake content can be noticed on much earlier events, such as the 2014 General Elections.

A recent published (RUEDIGER et al., 2017) by RUEDIGER et al. analyses the impact of bots on many recent events in Brazilian history. For instance, within all tweets related to the debate between Dilma Roussef and Aécio Neves, the presidential candidates who made it to the second round, over 10% of all tweets was artificially generated, to stimulate public opinion towards a certain Presidential candidate. Among Aécio Neves supporters the portion of interactions with automated accounts (bots being retweeted by other bots or regular accounts) reached 19.41%. In the discussions between profiles supporting Dilma, the amount was 9.76%.

This study also shows the progression of this phenomenon. After the impeachment of president Dilma Rousseff, the debate on labor reforms in the National Congress gained strength. This debate was impulsed by an apparent necessity of austerity measures to overcome the post-impeachment economic crisis, and this was seen by many politicians as an opportunity to modernise and reduce legislation on the matter. The opposition saw this as a decrease of workers rights and worsening of work conditions and social protection of the Brazilian State. This crisis erupted into the general strike on April 28, 2017, organized by labor unions and opposition, counting on a large turnout to convince the politicians and people alike of the dissatisfaction concerning these potential reforms. During the events of this General Strike, over 20% of all tweets discussing the subject were automated.

More recently during the 2018 General Elections, then candidate Jair Bolsonaro was accused many times of financing automated fake news messages over WhatsApp using funds outside of his legal campaign finances, prompting even a lawsuit on the Superior

Electoral Court investigating these claims - which is ongoing at the moment [1].

The impact of this influx of fake news in Brazilian politics has prompted action from the government, which has bolstered efforts to oppose fake news, with many events and seminars being promoted by non-governmental and governmental entities, like the International Seminar Fake News and Elections (Seminário Internacional Fake News e Eleições) [2], as well as calls for papers from the scientific community.

3.2 Fake news detection in portuguese

Even though there is a large recent interest in the field, automated fake news detection in portuguese is in its early stages. One of the reasons for this is that this problem is often solved using supervised learning algorithms, which require a labeled input dataset to operate on. However, up until recently, there were no available datasets in portuguese to be used to train the classifiers.

Recently, this has changed as (MONTEIRO et al., 2018) provided us with the first labeled fake news dataset in portuguese. The dataset, named Fakebr Corpus is available on github[3]. It contains 3,600 fake pieces of news, which were manually gathered and labeled. It is worth noting that they only kept in the final corpus news that were classified as totally fake, not keeping those that were only partially false. It also contains 3600 true pieces of news, which were collected in a semiautomatic way, by using a crawler to fetch news from reputable news sources and then manually checking the labels.

This work then tested the dataset, extracting many different features from the dataset and using those to provide a baseline classification using the LinearSVC implementation in Scikit-learn of the SVM algorithm. They achieved a maximum of 89% accuracy when taking into account all isolated features.

Other works have now been using this dataset to build their own solutions to the classification problem, such as (LEAL, 2018), which uses the same dataset with different types of neural networks to achieve an accuracy of 79% with the LSTM algorithm. Though there is already some promising work given the topic and context, there is still a lot to explore. There is still only one publicly available labeled dataset for fake news

[1] http://www.tse.jus.br/imprensa/noticias-tse/2018/Outubro/Corregedor-geral%20da%20Justica%20Eleitoral%20instaura%20acao%20da%20Coligacao%20O%20Povo%20Feliz%20de%20Novo%20c

[2] http://www.tse.jus.br/imprensa/noticias-tse/2019/Abril/seminario-internacional-fake-news-e-eleicoes-reunira-especialistas-nacionais-e-internacionais-em-brasilia

[3] https://github.com/roneysco/Fake.br-Corpus

classification in portuguese - and this dataset has been manually generated and labeled by the research team themselves. This limits the size of the dataset that can be created and also, unless the dataset is updated from time to time by the team, prevents us from having up-to-date news, which may adapt to have different deceptive cues from the ones currently in the dataset.

Also there are many approaches which have yet to be attempted in the context of news in portuguese. On top of that, the problem itself is still an open problem with no consensual solution.

4 IMPLEMENTATION AND METHODOLOGY

4.1 Dataset

4.1.1 Dataset Creation

At the time this work was started, the dataset provided by (MONTEIRO et al., 2018) was not yet available - and at the moment, to the best of our knowledge, it is the only dataset available with labeled fake news data in portuguese. As previously stated, this dataset was manually created, which comes with a series of drawbacks. Thus, one of the tasks in this work is to create a new, automatically generated dataset for fake news detection.

Fact checking is not a task that has been approached only from an automatic point of view. The recent relevance and impact of fake news has prompted many companies, governmental and non-governmental organizations to manually fact-check pieces of news and publish their review in order to warn the population and minimize the impact and propagation of the fake content. An example of one such entity is the EUFactCheck website [1], powered by the European Journalism Training Association (EJTA). There are also many such initiatives in Portuguese, such as the Fato ou Fake platform [2], created by the Globo group - the largest mass media group in Latin America.

These fact-checking initiatives are given a further degree of reliability by organizations which monitor and organize fact checking efforts worldwide, such as The International Fact-Checking Network (IFCN) [3]. They provide a code of principles [4] and a endorse fact-checking organizations that comply with their standards.

Therefore, there is an alternative to manually tagging the news ourselves. Instead, already assessed (labeled) claims can be retrieved from a series of fact checking websites, whose reliability has been endorsed either by organizations such as IFCN or by other trustworthy organizations. This is the approach we will follow in this work.

In order to create our dataset, we need to select some entities which will be the source of our pre-classified claims. The following sources for fact-checking data have been selected, based on the aforementioned reliability criteria:

[1] https://eufactcheck.eu
[2] https://g1.globo.com/fato-ou-fake/
[3] https://www.poynter.org/ifcn/
[4] https://ifcncodeofprinciples.poynter.org/

- Lupa: the first fact-checking agency in Brasil, in compliance with the requirements of the IFCN code of principles, an initiative by Grupo Folha [5]
- Aos Fatos: an independent initiative financed by many different organizations, also IFCN compliant [6]
- Publica: The first investigative journalism nonprofit agency in Brazil, another signatary of IFCN.
- E-farsas: The oldest fact-checking website in Brazil, maintained by the Record media group.

As neither of the selected sources provide an API to fetch the claim review information, we need to implement a program which will fetch the available data from their websites - commonly known as a web scraper (BOEING; WADDELL, 2017). This web scraper will have the tasks of first, on each website getting a list of all the claims that have been verified for each of the sources. Secondly, it needs to, for each of these claims, fetch the information which is relevant to the classification, and store it in a sufficiently structured way.

Each of these sources has different data available on them. The following information is available across all platforms, and will compose our dataset:

- Claim reviewed: The claim that was made by the piece of news which is being reviewed
- Claim review: The textual content of the review
- Date Published: The date this information was fact-checked
- Review URL: The URL where the Fact Checking information was retrieved from
- Review Rating: The conclusion reached by the fact checking. The values for this field vary between sources

The Review Rating has many different values across the platforms. Some classify news only between true or false, while others classify them as partially true, undefined, or other values. Thus, a normalization of these ratings is required. In Table 4.1 we show equivalence between the normalized values and original values.

As the structure of the web pages for each of these sources is different, a parser has been implemented for each of them, in order to fetch the information of interest, and the appropriate parser is used by the scraper when fetching data from each source.

[5] https://piaui.folha.uol.com.br/lupa/
[6] https://aosfatos.org

Table 4.1: Original values vs normalized values for each reviewed claim for each source

Source	*Original Value*	*Normalized Value*
aosfatos	verdadeiro	TRUE
aosfatos	falso	FALSE
efarsas	falso	FALSE
efarsas	indefinido	UNDEFINED
publica	verdadeiro	TRUE
publica	falso	FALSE
lupa	falso	FALSE
lupa	de olho	UNDEFINED
lupa	exagerado	UNDEFINED
lupa	verdadeiro	TRUE

The scraper has been implemented in python 3.6, in a collaborative work as a framework that was developed with a research group from the Federal University of Rio Grande do Sul, oriented by Prof. Dante Augusto Barone. The requests library was used for the required HTTP requests, the pandas library was used to help with the normalization and entity linking. A sequence diagram for the implemented scraper can be seen on figure 4.1 .

The generated Dataset contains over 2600 fact-checked claims, of which over 1100 were classified as either true or false, and will be used in the next step on our work. The other claims are kept on the dataset to enable future works that do not restrain themselves to a binary classification.The distribution of claims among each review rating is displayed on table 4.3, while Some examples can be found on Table 4.2.

4.1.1.1 Entity Linking

In order to enrich our dataset and enable it to be used for more classification approaches, and so that the news can be related to the entities that are affected by it, we have performed the task of entity linking (also known as named entity recognition and disambiguation) on the claim text. Named entity recognition, or entity extraction, is an important subtask of Information extraction, which involves identifying the names of all the people, organizations, and geographic locations in a text (GRISHMAN; SUNDHEIM, 1996). Entity linking goes a step further, taking these entities and verifying them with a pre-existing knowledge base (HACHEY et al., 2013).

To perform this task, a couple of text annotation tools that have entity linking capabilities imbued within them were evalated. (ROSALES-MéNDEZ; POBLETE; HOGAN, 2018) provides a quick comparison of such entity linking tools, a summation of which

Figure 4.1: Partial sequence diagram for implemented scraper

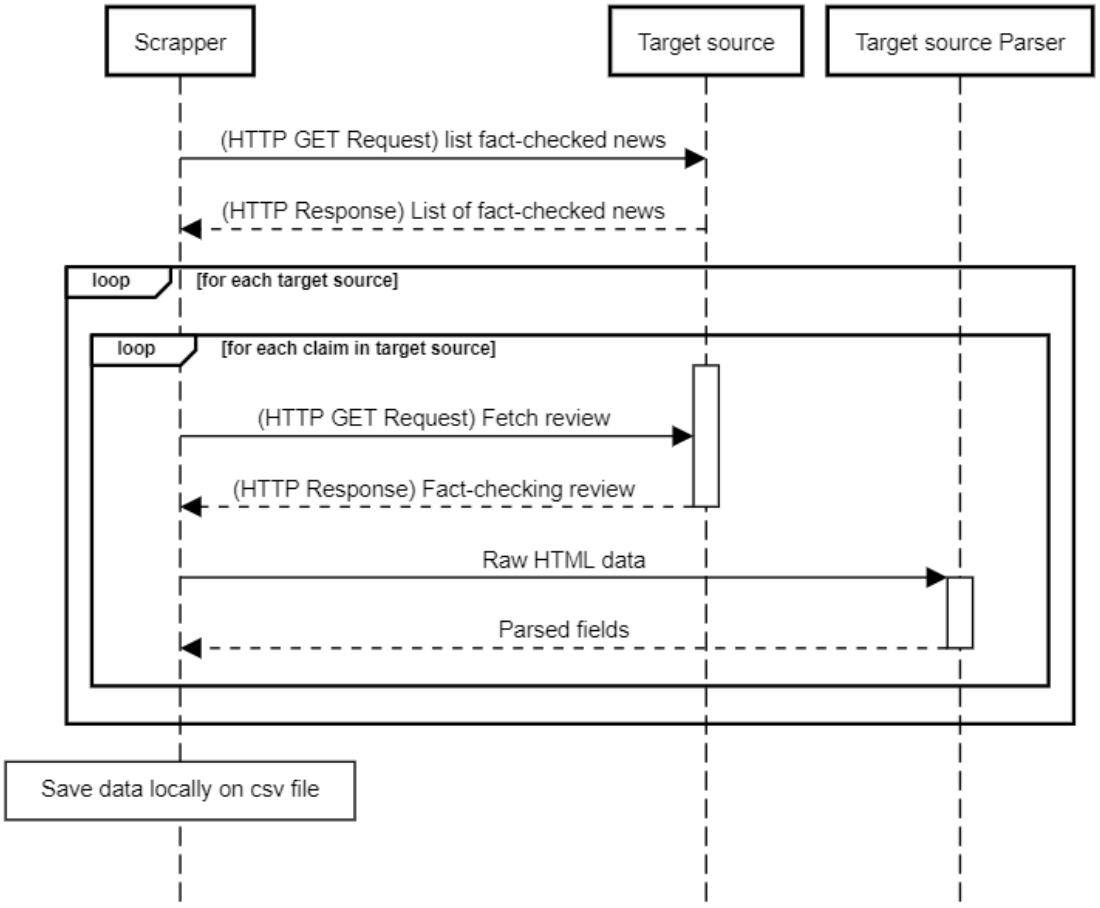


Table 4.2: Dataset samples

Claim reviewed	*Date Published*	*Review URL*	*Normalized Review Rating*
Tem aí uma questão de metodologia [do IBGE, sobre o número de desempregados]. Quem não procura emprego, é tido como empregado. Quem no ano passado trabalhou dois, três dias, é tido como empregado. Quem recebe o auxílio-desemprego é tido como empregado. – Jair Bolsonaro (PSL), em live realizada no Facebook do empresário Luciano Hang.	2018-10-26	https://apublica.org/2018/10/truco-em-economia-bolsonaro-cita-dados-falsos-e-haddad-subestima-e-acerta/	TRUE
Eu pedi a cassação do Temer. — programa eleitoral de Alvaro Dias (Podemos) em 4 de setembro de 2018 – Jair Bolsonaro (PSL), em live realizada no Facebook do empresário Luciano Hang.	2018-09-07	https://aosfatos.org//noticias/semana-1-os-erros-e-acertos-dos-presidenciaveis-na-propaganda-eleitoral-da-tv/	TRUE
OAB decide aprovar todos os candidatos que fariam a prova domingo [27/5]	2018-05-26	https://piaui.folha.uol.com.br/lupa/2018/05/26/verificamos-oab-prova-ordem/	TRUE
Um “lobisomem” foi filmado nos arredores da cidade de Guairaçá/PR?	2019-03-9	http://www.e-farsas.com/um-lobisomem-foi-filmado-nos-arredores-da-cidade-de-guairaca-pr.html	FALSE

can be seen on figure 4.2 . The Spotlight tool was selected as it was the only one which supported annotations in portuguese. Spotlight is powered by DBpedia, a wikimedia powered knowledge base (AUER et al., 2007).

Initially, using the web-available instance of DBpedia Spotlight was attempted, but the entity linking with spotlight was taking approx. 10 seconds per annotation, as it accessing was a publicly available shared instance. The solution to this was to run a local instance of Spotlight, with the code retrieved from [7] using an available model for portuguese entity extraction. Therefore, the final sequence diagram for the scraper can be seen on Figure 4.2.

[7]https://github.com/dbpedia-spotlight

Figure 4.2: Complete sequence diagram for implemented scraper

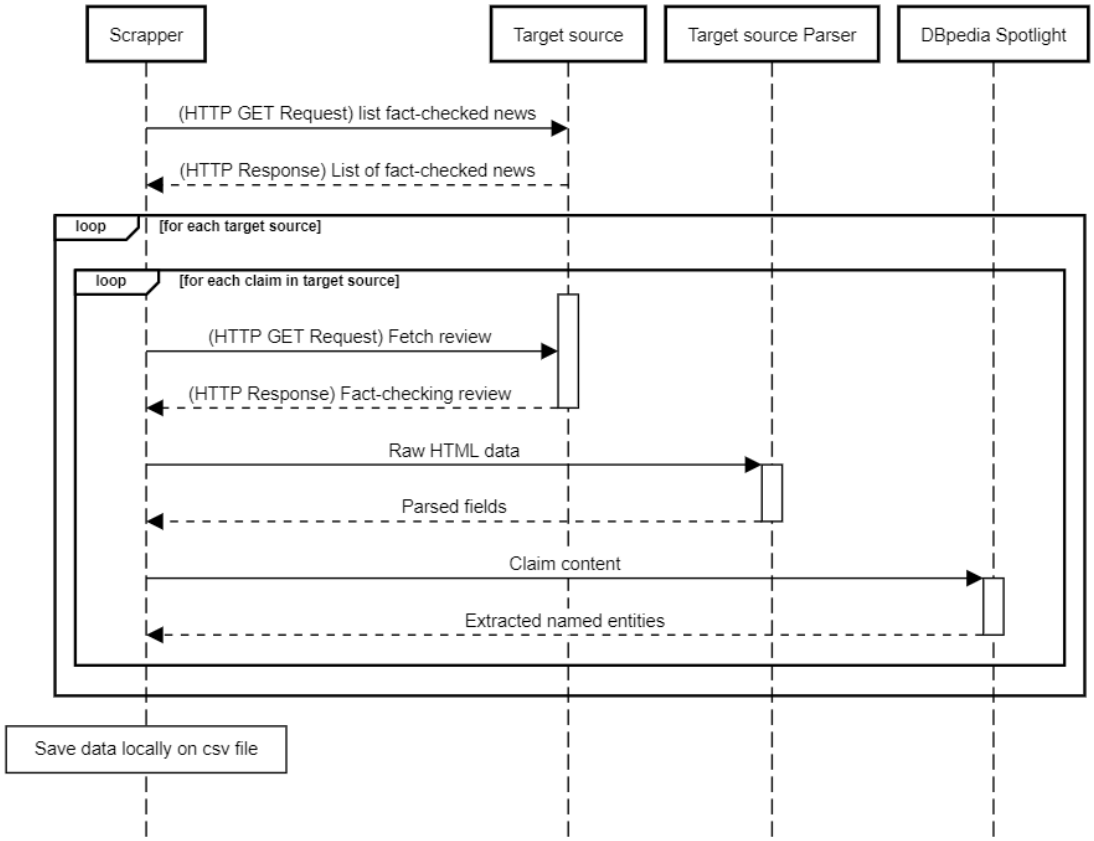


Table 4.3: Data distribution among normalized review ratings

TRUE	*FALSE*	*UNDEFINED*
558	557	1007

4.1.2 Data Quality

An important step is to analyse the quality of this dataset - How good is it?. (PIPINO; LEE; WANG, 2002) provides us with 16 data quality metrics, which can be subjectively analyzed. On the following list we briefly analyze our Dataset based on those metrics:

- Accessibility and security: As the dataset is publicly available on github[8], we can say that it is highly accessible and secure.
- Appropriate amount of Data: The data collected refers to a subset of the claims that have been analyzed by the previously defined sources. As datasets that have been used in the context of fake news detection range from 50 (Buzzfeed) up to 7200 pieces of news or more (FakeNewsNet), 1115 binary classified and 1123 news in total can be considered a reasonable amount of data.
- Believability and Objectivity: The credibility of the data, just as its impartiality, can be verified by the certifications that each of the fact-checking institutions selected as sources have, such as IFCN. Also, every entry on the dataset can be traced back to it's original entry on the source website by it's URL. As such, our data is highly believable.
- Completeness: Though there is no missing information on the dataset, there is more information which can be aggregated to it. For instance, the link to the original piece of news that was verified is not present on each claim, and neither are the engagements, which can be used for the classification. As such, we can say that the dataset has a moderate degree of completeness.
- Concise and consistent representation: The only information that can potentially be simplified are the extracted entities for each article. Aside from that, each and every field consists of data of the same type.
- Ease of manipulation: The data is provided as a single csv file, which is widely accepted by all data processing tools.
- Free-of-Error: This dataset can be said to be free of errors as the information is

[8]https://github.com/cristianormd/fakebr-claim-dataset

gathered in an automated fashion from the fact-checking source, all entries were checked for typing errors and sample entries were manually checked.

- Interoperability and Understandability: All entries are in the same language, which is the target language for this work (portuguese), and are strongly typed. Their meaning has been previously explained in this work.
- Relevancy: The relevancy of this dataset has been explained on previous sections.
- Reputation: There is no further work currently using the Dataset, so it is of low reputation : The automated tool can be triggered anytime to update the dataset with up-to-date news content.

4.2 Automated fake news detection

The objective of this section is to provide a baseline model for classification of fake news based on claims, using the created dataset.

As in the dataset we do not have any engagement information, we will implement only content based solutions. Also, we will focus on the data contained in the claim that has been reviewed by our reliable fact-checking sources. As we treat the fake news detection problem as a binary classification one, we will only be using the claims which have been classified as either TRUE or FALSE.

It also important to reiterate that our dataset contains the claim from the original news, not the original news themselves. As such, our models will be trained and validated using those claims - to the best of our knowledge, it is the first work to attempt to do so in the context of fake news.

All experiments were performed using the Jupyter Notebook (KLUYVER et al., 2016) tool for data science, using the Python programming language, version 3.6, running on an attached ipython kernel[9].

4.2.1 Feature selection and Engineering

As previously elaborated, many different features can be extracted from the text and used as input for the classification functions. An example of such feature are the entities we extracted and added into the dataset. (PÉREZ-ROSAS; MIHALCEA, 2015) provides

[9]https://ipython.org/

a review of features that have been used for fake news detection, and (MONTEIRO et al., 2018) validates them in the context of fake news in Portuguese. Based on those and the data we have at hand, the features that will be used are the following:

4.2.1.1 Bag of Words

A Bag of words is a simplifying representation used in natural language processing and information retrieval (IR). This representation transforms the original text in a set of words, with their number of occurrences in the text, disregarding their syntax, semantics and other grammatical information. Its usage for fake news has been highlighted in many of the already cited works, and has been analyzed in the context of text classification by (FÜRNKRANZ, 1998a).

Using this representation, instead of simply counting the frequency of each word, we will calculate its TF-IDF - short for term frequency–inverse document frequency. It is a numerical statistic which intends to represent how important each word is to the text (RAJARAMAN; ULLMAN, 2011).

This simple representation of the text usually undergoes further pre-processing steps such as lemmatization and stop word removal before being used to train the model, as they tend to better generalize the available information. In the context of this work, as we are doing mostly short text classification, some information may be lost on these generalizing and often error prone pre-processing steps (BOBICEV; SOKOLOVA, 2008). As such, we will be using the simple Bag of Words described here as well as a bag of words created from the pre-processed text as described in the next subsection.

4.2.1.2 Lemmatized Bag of Words

In this approach, we will be doing some pre-processing work on the text before adding each word to the bag, in an attempt to improve the quality of this feature for our classification task. For simplicity, we will refer to this approach as Lemmatized Bag of Words. The first step we perform is the stop-word removal from the claim. Stop words are words which do not add any value to the text, and are present in it merely to fulfill a certain necessary grammatical role. They tend to be the most common words in the language, and are generally filtered out before processing textual information. There is no universal list of stop words, therefore we have used the list of stop words for the Portuguese language provided by the NLTK (Natural Language Toolkit) python package. In this step we also

remove any non-alphabetical information.

The words that remain go through a lemmatization step. In a text, for instance, someone can use the words "walk" and "walked". In our TF/IDF representation, we would like those two words to be interpreted as the same entity - the cannonical form of the word, its lemma. Also, sometimes the same lemma may be used covering a different syntactical function (also called Part Of Speech) - such as in the phrases "I need to finish this delivery" and "I am delivering this". The lemmatization extracts the lemma from each word. It differs from the stemming procedure, which is also commonly used to achieve a similar result, in that it takes into account the syntactical function of the word being processed - whereas stemming does not (JIVANI et al., 2011) . We achieve this by using the spacy natural language library for python, which provides a pre-trained model for lemmatization in portuguese.

4.2.1.3 Shallow syntax: POS tags

Part of speech (POS) are categories of words which have similar grammatical properties. Those often mirror their syntactical function on the phrase. The task of POS tagging involves assigning each word to one of these POS categories (FÜRNKRANZ, 1998b). In this work, once again the capabilities of the spacy library are leveraged, using it's embedded pre-trained model for POS tagging in portuguese.

We have also attempted a classification by combining both the lemma of the words their respective tag into a single element, and using that as the feature for classification. We use the lemmatized word as any information added by the POS tag is already present in the original word.

4.2.1.4 Named Entities

We also attempt to classify the news based solely on the entities mentioned on the article, which are readily available in our dataset. (MOHAMMAD; SOBHANI; KIRITCHENKO, 2017) Has evaluated this approach when evaluating stances on tweets, and it has been used in the broader problem of deception detection. We attempt this approach to answer a hypothesis: Whether or not claims can be classified as fake or not based on the involved entities.

Finally, we will also be combining the extracted named entities with the other features, in order to validate if the features can complement themselves in the classification

step.

4.2.2 Model Validation

As already defined, supervised machine learning algorithms require an input test to train their classifier model. But how do we know if our model is good eanough? A common practice is to separate the dataset in two parts: A set of data which will be used as training data, and another set which will be used as validation data. This data is often randomly selected, such as to avoid biases generated by specific characteristics of the training set - this is often called sampling bias.

Even though the model is being trained with a randomly selected subset of data, it can still be biased. This is why we apply cross-validation techniques, to avoid such a bias. In particular, we will be using the k-fold cross-validation technique.

In this technique, we partition our entire input dataset S into k subsets. Turns are taken, and for each of the s_i subsets, the model is trained using s_i as its testing set, and $S = \{s_0, s_1 \ldots, s_k\}$ as the training set. This allows us to avoid the sampling bias (KOHAVI et al., 1995).

When partitioning our dataset, special attention was given such as to make sure that each fold had a similar amount of claims from each source, such as to avoid overfitting our model to be much better at classifying a certain style of claims from a certain source, and to avoid any bias that may be present on the way the claims are presented from each source - even though our sources are recognized as nonpartisan.

4.2.3 Evaluation metrics

There are many metrics which exist to measure the performance of binary classifiers. These metrics usually revolve on the amount of values on each of the quarters of the confusion matrix. The confusion matrix represents the 4 possible outcomes of the classification. In our case:

- True Positives (TP): Piece of news is fake and has been classified as fake
- False Positives (FP): Piece of news is fake, and has been classified as true
- False Negatives (FN): Piece of news is true, and has been classified as fake
- True Negative (TN): Piece of news is true, and has been classified as true

From those measurements, we derive the following metrics which will be used for evaluating our models:

- Precision: Represents the proportion of positive identifications that wherew actually correct. It is expressed as $\frac{TP}{TP+FP}$
- Recall: Represents the proportion of actual positives which were identified correctly. It is expressed as $\frac{TP}{TP+FN}$, in our scenario the number of news correctly classified as fake amongst those classified as fake
- F-score: It is a measure which takes on account both precision and recall. It is ideal when we assume false positives and false negatives have similar cost.

4.2.4 Algorithms

We will be using the following classifiers on our supervised learning:

4.2.4.1 Naive Bayes

Naive Bayes is a type of probabilistic classifier - a classifier which is able to predict a probability distribution of each feature over a set of classes (MARON, 1961).

This probabilistic classifier in particular applies the Bayes theorem to estimate the probability distributions. The definition of the Bayes theorem can be described mathematically as follows:

Bayes Theorem 1.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{4.1}$$

where A and B are events and $P(B) \neq 0$

$P(A \mid B)$*is a conditional probability: the likelihood of event A occurring given that B is true.*

$P(B \mid A)$ *is also a conditional probability: the likelihood of event B occurring given that A is true.*

$P(A)$ *and* $P(B)$ *are the probabilities of observing A and B independently of each other*

Given an instance of a problem $x = (x_1, \ldots, x_n)$, representing n features on our

input, our problem can be defined as

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)} \tag{4.2}$$

where C_k are the classes of the problem.

By using the chain rule, the problem can be rewritten as

$$p(C_k \mid x) = \frac{p(x_1 \mid C_k) \dots p(x_n \mid C_k)}{p(x_1) \dots p(x_n)} \tag{4.3}$$

Finally, as the denominator is constant across the dataset the probability distribution we want to learn is:

$$p(C_k \mid x) = Kp(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{4.4}$$

After the probability distribution is learned, the inputs are classified by calculating the probability of the given value to belong to each class, and using a maximum a posteriori rule to generate the output.

As our features represent frequencies of occurrences of words, an adaptation of Naive Bayes called Multinomial Naive Bayes has been used, which takes that into account when estimating the probability distribution function. This work uses the implementation provided by the scikitlearn[10] python library (RENNIE et al., 2003).

This algorithm was chosen as it is very simple, and Granik; Mesyura and others have achieved good results with this approach in the task of fake news detection in English.

4.2.4.2 Support Vector Machine - SVM

The other classifier used in this work uses a Support Vector Machine model. This model sees each input as a list of vectors in an n-dimensional space, represented in the form $\vec{v} = (\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$. This model attempts to create a function

$$f(\vec{v}) : \mathcal{X} \rightarrow C \tag{4.5}$$

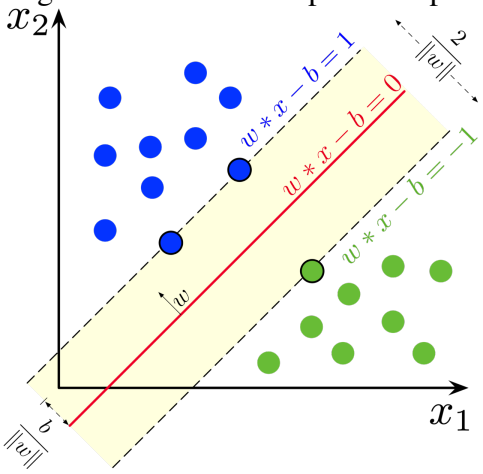
where C is the set of classes, and

$\mathcal{X}$ is the input space

with the maximum margin possible between the points on each class (Figure 4.3).

[10]https://scikit-learn.org/stable/

Figure 4.3: SVM - Simple Example



Maximum-margin hyperplane (in this case, a line) and for an SVM trained with samples from two classes. Samples on the margin are called the support vectors. Source: Wikimedia Foundation

Often, this is not possible to do achieve directly while trying to derive a linear function. Therefore, the Kernel trick is applied. This mathematical trick allows linear learning algorithms such as SVM to learn a nonlinear function or decision boundary. The kernel trick transforms the problem of approximating the function $f(\vec{v})$ by using another function, often called a kernel, that expresses each pair of inputs to the learning function as their inner product in a new vector space, such as that

$$k(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle_{\mathcal{V}} \tag{4.6}$$

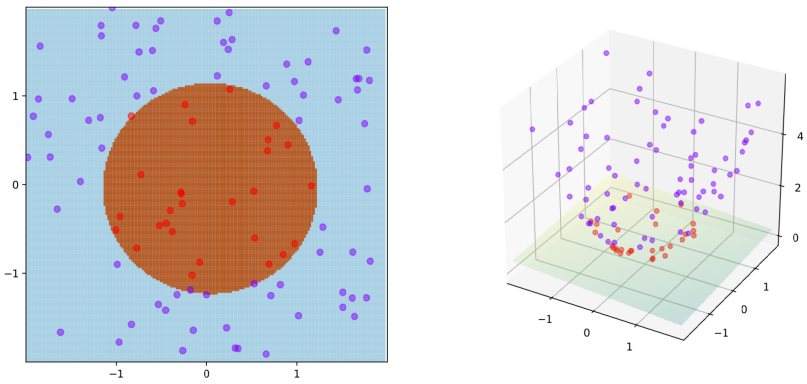
with φ being called a feature map, which transforms the inputs from input space $\mathcal{X}$ to a new input space V as follows.

$$\varphi \colon \mathcal{X} \to \mathcal{V} \tag{4.7}$$

Then, the function $f(\vec{v})$ can be approximated on the new vector space space V (Figure 4.4).

This approach has been widely used in the context of text classification, and more recently has been quite successful in the task of fake news detection (ZHOU; ZAFARANI, 2018), (MONTEIRO et al., 2018) . Once more, we use the implementation provided by scikitlearn. A linear kernel was experimentally verified to produce the best results.

Figure 4.4: SVM - Kernel Trick Example



An example of SVM with kernel given by $\varphi((a, b)) = (a, b, a_2+b_2)$. Source: Wikimedia Foundation

5 RESULTS AND ANALYSIS

Table 5.1: Classification results for each algorithm and feature set, using k-fold cross-validation with k=5

Model	**Metrics**		
	Precision	*Recall*	$F-score$
Naive Bayes			
Bag of words	**0.687**	**0.674**	**0.669**
Lemmatized Bag of words	0.648	0.636	0.630
POS tags	0.560	0.541	0.509
Lemmatized Bag of words + POS tags	0.658	0.647	0.643
Named Entities	0.659	0.645	0.639
Bag of Words + Named Entity	0.674	0.662	0.658
Lemmatized BOW + POS + Named Entity	0.687	0.672	0.668
Support Vector Machine			
Bag of words	0.657	0.656	0.656
Lemmatized Bag of words	0.646	0.645	0.645
POS tags	0.553	0.536	0.518
Lemmatized Bag of words + POS tags	0.649	0.648	0.648
Named Entities	0.627	0.626	0.625
Bag of Words + Named Entity	0.656	0.653	0.652
Lemmatized BOW + POS + Named Entity	**0.661**	**0.659**	**0.659**

A pair of results that distinct themselves from the rest are the models created using only the POS tags as a classification feature. All their performance metrics - and in particular F-score - held results really close to 0.5, therefore providing a result which is not significantly better than chance.

Apart from those results, all trained models we had an F-score of over 0.62. This margin considerably above chance indicates that the deceptive cues are present in the claims of a piece of news, and that by using only a piece of news claims we are able to - albeit not with a high f-score - detect deceitful news.

Surprisingly, the best result accross all metrics was achieved with the simple Bag of Words feature on the model trained using the Naive Bayes algorithm. This corroborates our hypothesis that, for texts with small textual content, pre-processing the text may lead to the loss of relevant information.

A model with an extrememly smilar performance was the one that took into

account all of the features, using the pre-processed bag of words with the POS tags and the Named Entities. It is not so surprising that by using all features we were able to achieve good reasults, but this indicates that the Named Entities indeed, as hypothesised, are a distinctive feature of fake news which improved the classification performance metrics when compared with the emmatized Bag of words + POS tags. This is further emphasized by the fact that, amongst all the models trained using the Support Vector Machine algorithm, using all features provided the best results.

Furthermore, using Named Entities alone as a feature for fake news detection enabled us to achieve over 0.62 f-score with SVM and over 0.63 with Naive Bayes. Which not only reinforces our conclusion, but points that Named Entities are a feature of similar importance to others in recognizing deceptive cues.

When comparing the algorithms used, we can clearly see that Naive Bayes offered a best average performance accross all metrics when compared to SVM. This was surprising as most of the literature, (ex: (MONTEIRO et al., 2018)) tends to prefer SVM over other similar classification algorithms.

Finally, it is clear to see that excepting the POS tags outliers the results for all of the trained models are quite similar in performance, accross all metrics. It may be the case that some of the inputs are more easily classifiable, whereas others are not - and those are the cases in which all trained models are able to achieve a successful classification, or news which are easily identifiable by a certain feature, may not be identified by another. A more in-depth analysis of the generated models would be required to better isolate the cause.

6 CONCLUSION

This work proposes a new dataset for the problem of fake news detection in portuguese, to the best of my knowledge the first one to not be tagged manually by the research team, and to use sources whose unpartisanship and unbiased review are endorsed by well renowned international organizations, such as IFCN.

It also provides the results of different classification models, using different algorithms and features, which can be used as basis for future work using this or other datasets. Some of those models are the first fake-news classification attempts which use named entities as one of the features for fake news classification, which showed high promise as a solution for the problem, helping us reach one of our best results.

Finally, the contributions of this work to the recent and growing research topic of automated fake news detection in Portuguese aim to foster the development of the field and pursue new approaches to the solution of this problem. It establishes a base for future work, in hopes that we can all move forward towards automated fake news detection in Portuguese.

6.0.1 Future Work

Currently, our dataset contains only the claims and the reviews of each claim made by the original piece of news. A next step towards building an unbiased fake news dataset would be to, based on the claim and the information provided by the fact-checking organization that has made the review, search the web for the original pieces of news that prompted that claim verification, and add those to the dataset. Given that, one could also search for the engagements and reference to that particular piece of news, to enable this dataset to support more approaches of fake news detection.

In terms of the baseline classification that was made, more algorithms could be compared in the context of fake news detection in portuguese - there are few works that have explored this aspect. Also, there is at the moment no comparison between the similarities and differences of the problem of fake news detection between portuguese and other languages - to answer questions such as: are there any algorithms or features that are better suited for classification in portuguese in particular? Is there an algorithm or set of features which are universally better, regardless of the language?

Lastly, the news classified as UNDEFINED could be used on future work which

does not constrain the problem of fake news detection into a binary classification problem, but rather attempt to classify news which are "half truths" into a spectrum of veracity.

REFERENCES

AHMED, H.; TRAORE, I.; SAAD, S. Detection of online fake news using n-gram analysis and machine learning techniques. In: . [S.l.: s.n.], 2017. p. 127–138. ISBN 978-3-319-69154-1.

ALLCOTT, H.; GENTZKOW, M. 2017.

AUER, S. et al. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. [S.l.]: Springer, 2007. p. 722–735.

AYODELE, T. O. Types of machine learning algorithms. In: **New advances in machine learning**. [S.l.]: IntechOpen, 2010.

BOBICEV, V.; SOKOLOVA, M. An effective and robust method for short text classification. In: **AAAI**. [S.l.: s.n.], 2008. p. 1444–1445.

BOEING, G.; WADDELL, P. New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings. **Journal of Planning Education and Research**, SAGE Publications Sage CA: Los Angeles, CA, v. 37, n. 4, p. 457–476, 2017.

BOND R.M., F. C. J. J. K. A. M. C. S. l. J.; FOWLER, J. The science of fake news. **Science**, v. 359(6380), p. 1094–1096, 2018.

CHEN, Y.; CONROY, N. J.; RUBIN, V. L. Misleading online content: Recognizing clickbait as false news. In: ACM. **Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection**. [S.l.], 2015. p. 15–19.

CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. **Proceedings of the Association for Information Science and Technology**, Wiley Online Library, v. 52, n. 1, p. 1–4, 2015.

CORNER, J. Fake news, post-truth and media–political change. **Media, Culture & Society, 39(7)**, 2017.

DIFRANZO, D.; GLORIA-GARCIA, K. Filter bubbles and fake news. **XRDS: Crossroads, The ACM Magazine for Students**, v. 23, p. 32–35, 04 2017.

FENG, S.; BANERJEE, R.; CHOI, Y. Syntactic stylometry for deception detection. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2**. [S.l.], 2012. p. 171–175.

FENG, V. W.; HIRST, G. Detecting deceptive opinions with profile compatibility. In: **Proceedings of the Sixth International Joint Conference on Natural Language Processing**. [S.l.: s.n.], 2013. p. 338–346.

FENG, V. W.; HIRST, G. Detecting deceptive opinions with profile compatibility. In: **Proceedings of the Sixth International Joint Conference on Natural Language Processing**. [S.l.: s.n.], 2013. p. 338–346.

FÜRNKRANZ, J. A study using n-gram features for text categorization. **Austrian Research Institute for Artifical Intelligence**, v. 3, n. 1998, p. 1–10, 1998.

FÜRNKRANZ, J. A study using n-gram features for text categorization. **Austrian Research Institute for Artifical Intelligence**, v. 3, n. 1998, p. 1–10, 1998.

FUSSELL, P. **The Great War and Modern Memory**. [S.l.]: Oxford University Press US, 2000. ISBN ISBN 0-19-513332-3.

GELFERT, A. Fake news: A definition. **Informal Logic 38.1 (2018) 84-117**, 2017.

Granik, M.; Mesyura, V. Fake news detection using naive bayes classifier. In: **2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)**. [S.l.: s.n.], 2017. p. 900–903.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**. [S.l.: s.n.], 1996. v. 1.

GUESS A., N. B. . R. J. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. **European Research Council 9**, 2018.

GUPTA, A. et al. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: ACM. **Proceedings of the 22nd international conference on World Wide Web**. [S.l.], 2013. p. 729–736.

HACHEY, B. et al. Evaluating entity linking with wikipedia. **Artificial intelligence**, Elsevier, v. 194, p. 130–150, 2013.

HAYS, J.; EFROS, A. A. Scene completion using millions of photographs. **ACM Transactions on Graphics (TOG)**, ACM, v. 26, n. 3, p. 4, 2007.

HUH, M. et al. Fighting fake news: Image splice detection via learned self-consistency. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 101–117.

JIN, Z. et al. Novel visual and statistical image features for microblogs news verification. **IEEE transactions on multimedia**, IEEE, v. 19, n. 3, p. 598–608, 2016.

JIVANI, A. G. et al. A comparative study of stemming algorithms. **Int. J. Comp. Tech. Appl**, v. 2, n. 6, p. 1930–1938, 2011.

KHAN, A. et al. A review of machine learning algorithms for text-documents classification. **Journal of advances in information technology**, Academy Publisher, PO Box 40 Oulu 90571 Finland, v. 1, n. 1, p. 4–20, 2010.

KLUYVER, T. et al. Jupyter notebooks-a publishing format for reproducible computational workflows. In: **ELPUB**. [S.l.: s.n.], 2016. p. 87–90.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

KONSTANTINOVSKIY, L. et al. Towards automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection. **arXiv preprint arXiv:1809.08193**, 2018.

Kwon, S. et al. Prominent features of rumor propagation in online social media. In: **2013 IEEE 13th International Conference on Data Mining**. [S.l.: s.n.], 2013. p. 1103–1108. ISSN 1550-4786.

LAZER, D. M. et al. The science of fake news. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018.

LEAL, I. H. D. S. O uso de aprendizagem de máquina para identificação e classificação de fake news no twitter referentes a eleição presidencial de 2018. 2018.

LIU, Y.; WU, Y.-F. B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: **Thirty-Second AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2018.

MANGASARIAN, O. L.; MUSICANT, D. R. Lagrangian support vector machines. **J. Mach. Learn. Res.**, JMLR.org, v. 1, p. 161–177, sep. 2001. ISSN 1532-4435. Available from Internet: <https://doi.org/10.1162/15324430152748218>.

MARON, M. E. Automatic indexing: an experimental inquiry. **Journal of the ACM (JACM)**, ACM, v. 8, n. 3, p. 404–417, 1961.

MIGNE, J.-P. **Patrologia Latina. Volume 143**. [S.l.]: Routledge, 1891.

MOHAMMAD, S. M.; SOBHANI, P.; KIRITCHENKO, S. Stance and sentiment in tweets. **ACM Transactions on Internet Technology (TOIT)**, ACM, v. 17, n. 3, p. 26, 2017.

MONTEIRO, R. A. et al. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 324–334.

NEWMAN, M. L. et al. Lying words: Predicting deception from linguistic styles. **Personality and social psychology bulletin**, Sage Publications, v. 29, n. 5, p. 665–675, 2003.

PAPANASTASIOU, Y. Fake news propagation and detection: A sequential model. **Available at SSRN 3028354**, 2018.

PÉREZ-ROSAS, V.; MIHALCEA, R. Experiments in open domain deception detection. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2015. p. 1120–1125.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, ACM, v. 45, n. 4, p. 211–218, 2002.

POMERANTSEV, P. cited in 'the post-truth world: Yes, i'd lie to you'. **The Economist**, v. 10, 2016.

RAJARAMAN, A.; ULLMAN, J. D. Data mining. In: ____. **Mining of Massive Datasets**. [S.l.]: Cambridge University Press, 2011. p. 1–17.

RENNIE, J. et al. Tackling the poor assumptions of naive bayes classifiers (pdf). In: ICML. [S.l.], 2003.

ROSALES-MéNDEZ, H.; POBLETE, B.; HOGAN, A. What should entity linking link? In: . [S.l.: s.n.], 2018.

RUEDIGER, M. A. et al. Robôs, redes sociais e política no brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018. FGV DAPP, 2017.

RUSSELL, B. **A History of Western Philosophy**. [S.l.]: Routledge, 2004.

SHU, K. et al. Fake news detection on social media: A data mining perspective. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 19, n. 1, p. 22–36, sep. 2017. ISSN 1931-0145. Available from Internet: <http://doi.acm.org/10.1145/3137597.3137600>.

SHU, K.; WANG, S.; LIU, H. Exploiting tri-relationship for fake news detection. **arXiv preprint arXiv:1712.07709**, 2017.

SILVERMAN, C. et al. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. **Buzzfeed News**, v. 20, 2016.

SUWAJANAKORN, S.; SEITZ, S. M.; KEMELMACHER-SHLIZERMAN, I. Synthesizing obama: learning lip sync from audio. **ACM Transactions on Graphics (TOG)**, ACM, v. 36, n. 4, p. 95, 2017.

TSAI, Y.-H. et al. Deep image harmonization. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 3789–3797.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018. ISSN 0036-8075. Available from Internet: <https://science.sciencemag.org/content/359/6380/1146>.

ZHOU, X.; ZAFARANI, R. Fake news: A survey of research, detection methods, and opportunities. **CoRR**, abs/1812.00315, 2018. Available from Internet: <http://arxiv.org/abs/1812.00315>.