University of Wolverhampton

Dissertation 7LN007/UM1

Is this really fake?
A study of Spanish and French Automatic Fake News Detection

Isabel Olmos Canovas

Supervisors

Dr. Hadeel Saadany & Dr. Rocio de Caro

July 5, 2022

Dedication

I dedicate my dissertation project to my family and life partner. A special feeling of gratitude to my loving parents, Antonio and Isabel whose words have always encouraged me to pursue my dreams. I also dedicate this dissertation to my grandfather; although he is no longer of this world, his memories continue to regulate my life.

Acknowledgement

My accomplishment of this thesis is a result of support and guidance from many people around me. I would like to offer my special thanks to Dr. Hadeel, who has been an ideal mentor and thesis supervisor. Her dedication and effort have been instrumental in guiding me throughout the thesis. I would also like to especially thank Dr. Rocío, for her invaluable support with the linguistic analysis aspects. I am grateful to both of them for providing me this wonderful opportunity to pursue this thesis under their supervision and amidst their busy schedule for providing me insightful and valuable feedback. I would also love to express my earnest gratitude to all my teachers at the University of Wolverhampton. I am deeply grateful for all you have taught me, and for opening up new horizons of linguistic research. Finally, I would like to acknowledge with gratitude, the support of my parents, my brother, my sister, and my two nieces, for always being my pillars of strength and showing huge trust and confidence in me.

Abstract

Fake news is widely regarded as one of the major hazards to global commerce, journalism, and democracy. Recent political events have led to an increase in the popularity and propagation of fake news articles. With the widespread dissemination of information via digital media platforms, it is of utmost importance for individuals and societies to be able to judge the credibility of any piece of news. Since humans are inconsistent, if not outright poor detectors of fake news, efforts have been made to automate the process of fake news detection. The problem of automatic Fake News detection has been approached in this study from Natural Language Processing (NLP) and Machine Learning (ML) perspectives. Although in NLP research there are differing opinions on what may fall under the category of fake news, yet generally speaking any story, news, or headlines that are deliberately created to misinform or mislead the reader is considered fake news. This project is aimed at performing various analyses to identify the linguistic properties that are predominantly present in deliberate fake news. Its contribution is twofold: it builds a number of machine learning models including a robust deep neural network capable of capturing fake news, and statistically identifies a number of linguistic features peculiar to Spanish and French deliberate fake news. The study will be based on the combination of two Spanish datasets, comprising a total of 2,971 articles, and two French datasets downloaded from Github users, containing 27,816 pieces of news. The selected datasets are concerned with different topics, such as: politics, education, sport, and the environment. Performances of the different models implemented are compared in terms of accuracy, precision, recall, and F1 score. The results of this project demonstrate the ability for machine learning to be useful in this task with over 98% accuracy.

Contents

List of Figures

List of Tables

Introduction and background

Deception detection consists of investigating practices used to determine an individual's truthfulness and credibility; it involves activities such as the identification of scientific fraud, false tweets, or fake pieces of news. The term 'Fake News' rose to popularity during the 2016 US Presidential Election campaign, where the term was increasingly weaponized by both the American right and left to discredit and denigrate the political opponent. Disseminators of fake news frequently have a vested interest, such as harming a person's reputation, threatening public safety, or profiting from false allegations. According to (Balmas, 2014), misleading information rises before significant political events, as a result of fake news leading to feelings of inefficacy, alienation, and cynicism towards particular political candidates. An instance of a fake news campaign that illustrates the massive impact of fake news includes a sudden shortage of salt in Chinese shops following a false allegation that iodized salt would help prevent the effects of radiation following the Fukushima nuclear disaster in Japan. Another clear example worth illustrating would be the $130 billion loss in the USA stock market (2013) as a result of a false new claim that US President Barack Obama had been harmed in an explosion (Rapoza, 2017).

Recent literature reports on cases of social bots imitating humans to manipulate discussions, alter the popularity of users, pollute content, spread misinformation, hoaxes and even perform terrorist propaganda and recruitment actions. BotOrNot[1], introduced by Davis et al. (2016), is a publicly available online service that automatically evaluates if a Twitter account is a real account or a social bot. This system uses over a thousand features to determine the degree to which a Twitter account resembles the recognized characteristics of social bots. Another tactic that promotes the dissemination of false pieces of news is 'Clickbait', a publicity tactic that uses sensitive headlines intended to attract consumers' attention and drive click-throughs to the publisher's website. Sensational news articles or headlines are typically utilized to click the user's browsing into ads. More clicks on the ad entail a higher profit (Chen et al., 2015).

Considering the fact that online news sources are not frequently verified, its content cannot be entirely trusted as if emanated from a reliable source. The vast volume of material disseminated and the pace with which it spreads on social media sites, poses a challenge in determining its veracity in a timely way. As a result, detecting fraudulent news articles has emerged as a prominent research topic in Natural Language Process-

[1] https://botometer.osome.iu.edu/

ing (NLP). Specifically, automatic detection approaches based on Artificial Intelligence (AI) and Machine Learning (ML) have been studied to counteract the proliferation and propagation of misleading information by developing computer algorithms that can assess the trustworthiness of online materials (Pérez-Rosas et al., 2017). Research on fake news is directed towards preventing the spread of misinformation over the internet by distinguishing fake news from real news. This task is essentially handled as a text classification problem. Nevertheless, due to several circumstances, detecting fake news on social media is extremely complicated. First of all, collecting misleading pieces of news and manually labelling them is a laborious process. Consequently, some online news datasets only include a reduced number of samples, which is insufficient for training a generic model for application. Moreover, when writing fraudulent news, human beings tend to use words strategically to avoid being discovered (Yang et al., 2018), which complicates the detection process.

Given the scarce research of fake news detection in languages other than English, this thesis puts forward the hypothesis that Spanish and French fake news can be automatically identified via computational tools. It proposes a number of baseline and hierarchical neural network models which are capable of automatically identifying misleading content. The experimental findings on the Spanish and French news datasets analysed in this thesis, prove the usefulness of employing word embeddings as input to Recurrent and Convolutional Neural Networks in detecting fake news with an accuracy of up to 98.7%. This study will demonstrate that semantic information provided by word-embedding vectors is sufficient for the effective detection of fake news in two different languages, which confirms the ability of computer algorithms to automatically detect fake news.

To understand the research background, this project first presents a literature review of fake news detection where different proposed models and research findings are compared. The literature review chapter will assess and contrast the different proposed models in NLP research for automatic detection of fake news in English, French and Spanish. Based on previous studies, the researcher opted to follow a distinct methodology in chapter two. The alternative steps followed in the experimental process are described in Chapter 2: corpus selection and description, data cleaning and pre-processing, extraction of linguistic features, and implementation of traditional and neural network-based algorithms. Chapter 3 presents the results and evaluation of the findings obtained in this project. Additionally, chapter four will include a brief discussion of different linguistic aspects such as error analysis. The final chapter will highlight some concluding remarks on the project including the strengths and limitations of the proposed study as well as potential further research.

Chapter 1

Literature Review

Despite all the research conducted to date, the task of fake news detection remains extremely complex; not only for humans, but also for machines. According to Ruchansky et al. (2017), humans are not particularly skilled at differentiating between genuine and misleading news. In their study, 75% of individuals classified fake news as accurate news; as a result of fake news resembling real news. Similarly, an experiment performed by Ribeiro et al. (2017), demonstrated that humans tend to classify news articles that they do not agree with as fake news. In an attempt to discover the basic motive why humans tend to believe fake news, Dwyer (2019) identified seven reasons: confirmation bias (favoring of information that confirms existing beliefs), lack of credibility evaluation (need for critical thinking), attention and impatience (incapability to read the whole article), cognitive laziness (when human-beings fail to engage evaluation and reflective judgment), targeted emotions (involving irrational thinking), the 'illusory truth effect' (the more exposure to certain information, the more likely it is to believe that information), and social pressure. Therefore, due to the readers' uncertainty and biased perception towards fake news recognition, a necessity for approved news from reliable sources and fact-checking websites has arisen.

In fact-checking websites, experts and journalists manually verify assertions against evidence based on previously stated facts. These websites are accessible in a variety of languages globally, where professional editors manually scrutinize the truthfulness of news. Some examples of online fact-checking resources are: FactCheck.org[1], Factmata.com[2], TruthOrFiction.com[3], OpenSecrets.org[4], PolitiFact[5], Snopes[6], Channel4.com[7], BS detector[8] and GossipCop[9]. Figure 1.1 illustrates two samples of manual fact-checking websites. The main limitation of fact-checking websites is the time delay

[1] https://www.factcheck.org/

[2] https://factmata.com/

[3] https://www.truthorfiction.com/

[4] https://www.opensecrets.org/

[5] https://www.politifact.com/

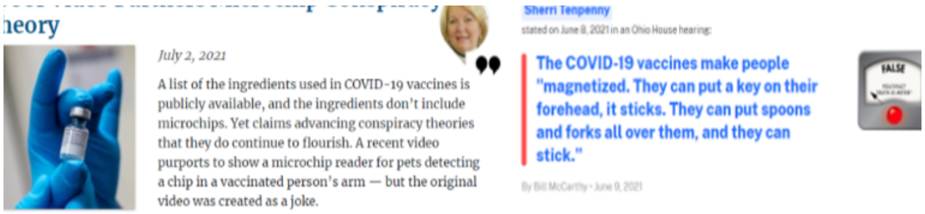
[6] https://www.snopes.com/

[7] https://www.channel4.com/news/topic/fake-news

[8] https://www.bsdetector.info/

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

involved in manual verification. It must also be noted that the vast majority of these resources consist of very concise statements that are primarily based on the verification of political news. Therefore, the practical applicability of those systems becomes restricted, due to the significant range of news types and formats, as well as the rapid dissemination of deceptive information in the social network. Consequently, it might be that by the time a piece of news is manually verified, it would have already been published on different platforms.



heory

July 2, 2021

A list of the ingredients used in COVID-19 vaccines is publicly available, and the ingredients don't include microchips. Yet claims advancing conspiracy theories that they do continue to flourish. A recent video purports to show a microchip reader for pets detecting a chip in a vaccinated person's arm — but the original video was created as a joke.

(a) Factcheck.org

Sherri Tenpenny

stated on June 8, 2021 in an Ohio House hearing:

The COVID-19 vaccines make people "magnetized. They can put a key on their forehead, it sticks. They can put spoons and forks all over them, and they can stick."

By Bill McCarthy • June 9, 2021

(b) PolitiFact

Figure 1.1: Illustrations of Manual Fact-checking Websites.

The growth of the Internet community, and the rapid spread of online information, have resulted in automatic fake news detection attracting the interest of the AI (Artificial Intelligence) research community. Hence, Automatic fake news detection is aimed at decreasing the amount of time and effort required by humans to detect fake news and assist individuals in preventing its spread. With advancements in computer science subfields such as Machine Learning, Data Mining (DM), and Natural Language Processing, have resulted in the detection of misleading news being addressed from a variety of viewpoints (Oshikawa et al., 2018). In this context, the partnership of academics and journalists is becoming increasingly important in the development of automated fake news detection systems (Özgöbek and Gulla, 2017). Such systems have previously been offered to protect users against internet disinformation. Twitter, for instance, integrates the Twitter Crawler, a component built on a NER algorithm that collects tweets in a database (Atodiresei et al., 2018). Hence, when the user wants to verify the accuracy of the news articles, they can copy the link directly into the application.

Given the importance of automatic fake news detection, new research initiatives have arisen in the European Union. SocialTruth[10], for instance, is one of the most paramount European Union research initiatives, which is funded within Horizon 2020 (H2020) program, the largest Research and Innovation Framework Programme in the history of the European Union. This project aims at detecting fake news by both professionals (e.g., journalists) and individuals (daily social media users), using lifelong Learning Machines that constantly accumulate experience and learn new paradigms of fake news. Other wider projects have been launched, such as SOMA[11] (Social Observatory for Disinformation and Social Media Analysis), which provides support to a European community that will jointly tackle disinformation (Giełczyk et al., 2019). Recently, SOMA has conducted an investigation on a series of hoaxes and conspiracy

[10] http://www.socialtruth.eu/

[11] https://www.disinfobservatory.org/

theories related to an alleged correlation between Covid-19 and 5G technology. These hoaxes on 5G and coronavirus can be summarized in two main streams: either the new coronavirus is activated by 5G, or Covid-19 does not exist and the symptoms people are experiencing are reactions to 5G waves. This investigation revealed that disinformation on this issue is leading to unexpected and dangerous consequences including criminal activities.

The task of fake news detection has been studied from different perspectives and sub-areas of computer science, nevertheless, the focus of this literature survey will be on the NLP approach. This section will provide deep insight into fake news and recent research for its automatic detection, emerging challenges, solutions proposed, as well as the advantages and limitations of each learning solution. This literature survey will be divided into five sections: (a) a brief explanation of fake news' categorization and characterization, (b) a description of the most prominent existing datasets in English, (c) insight of the state-of-the-art research efforts for automatically classifying fake news in English, (d) NLP studies of Spanish and French fake news detection, and (e) concluding remarks on this literature review.

1 Fake news: definition, categorization and characterization

There have been many discussions about the definition and meaning of the term 'Fake News'. To understand it correctly, it is first necessary to refer to the First Draft News[12], a project aimed to fight false information online, which was founded in 2015 by nine organizations, including Facebook, Twitter, the Open Society Foundations and several philanthropic organizations. The First Draft Project established three overarching types of information disorder, which are shown in Figure 1.2: misinformation, disinformation and malinformation. Misinformation is defined as false content that is unintentionally or unknowingly disseminated. On the other hand, disinformation implies the presence of false content that is intentionally disseminated with an intent to cause harm. The third term 'malinformation' comprises genuine information that is shared with the intent to harm.

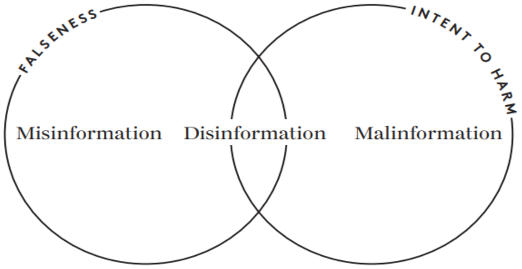


Figure 1.2: Types of information disorders.

[12] https://firstdraftnews.org/

Within these three overarching types of information disorder, First Draft News also refers to seven categories in an attempt to help people understand the complexity of this ecosystem. This classification is based on the information's content, creator's motivations and dissemination methods. The distinct categories are included in Figure 1.3: (a) Satire or parody: no intention to cause harm, but has potential to deceive, (b) False connection: when content is not supported by headlines, visuals, or captions, (c) Misleading content: deceptive use of information to frame an issue or person, (d) False context: when real content is combined with fabricated contextual information, (e) Imposter content: when truthful information is impersonated, (f) Manipulated content: when factual or imagery information is manipulated to deceive, (g) Fabricated content: new content that is entirely fake, and generated to deceive and cause harm.

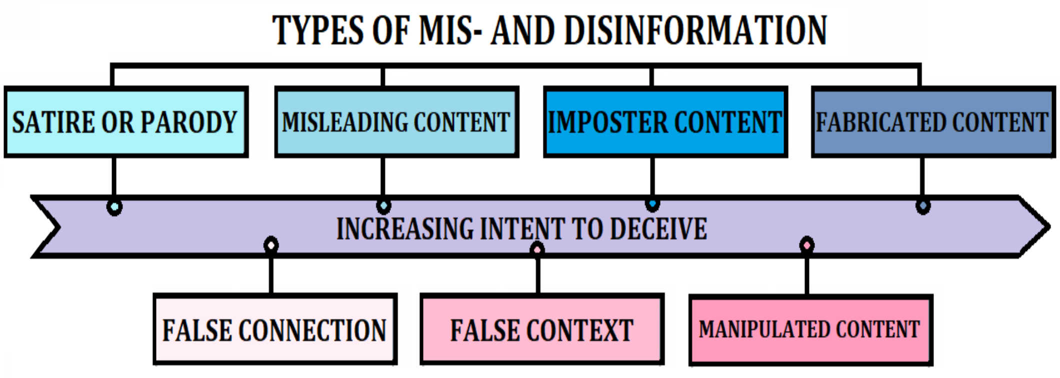


Figure 1.3: Types of misinformation and disinformation.

Different from misinformation and disinformation, the universe of "fake news" is much larger than simply false news stories. Fake News is described as any purposefully crafted, emotionally charged, sensational, deceptive or even completely fabricated content that imitates the form of mainstream news (Zimdars and McLeod, 2020). That is to say, any news stories that have been fabricated, with no verifiable facts, sources or quotes. News-related theories highlight the existence of discriminatory features that can be employed to automatically differentiate between fake and legitimate news, as can be seen in Zhou and Zafarani (2020). The main theories described by the authors are: Undeutsch hypothesis (Undeutsch, 1967), Four-factor theory (Zuckerman et al., 1981), and reality monitoring (Johnson and Raye, 1981). The Undeutsch hypothesis, claims that a statement based on a factual experience differs in content and quality from that of fantasy. The four-factor theory asserts that lies are expressed differently in terms of arousal (liars may experience greater undifferentiated arousal than truth-tellers, which can be detected via speech errors, hesitations, repetitions), behaviour control (over-control of body language), emotion (emotions change when lying), and thinking from truth (this leads liars to take longer in speaking with more pauses and using more generalities to avoid getting trapped by specific detail). On the other hand, reality monitoring theory defends the idea that actual events are characterized by higher levels of sensory-perceptual information.

These hypotheses and fundamental theories have motivated the implementation of style-based deception studies. It should also be noted that these theories initially targeted deceptive statements such as misinformation. Nonetheless, because fake news is a related concept, it has been successfully applied by previous authors, such as Zhou and Zafarani (2020). Based on the already mentioned theories, Zhou and Zafarani (2020) grouped some content and stylistic features along ten parallel dimensions: quantity (i.e., word or character counts), complexity (i.e., the average number of clauses per sentence), uncertainty (percentage of modal verbs), subjectivity (percentage of subjective verbs), non-immediacy (group reference: 1st person plural pronouns), sentiment (sentiment polarity), diversity (i.e., redundancy), informality (typographical error ratio), specificity (sensory ratio), and readability (Flesch-Kincaid and Gunning-Fog index).

According to fake news detection datasets in NLP research, fake news can be broadly categorized into three types: a) satire: mimics legitimate news but still gives the reader cues that it should not be taken seriously, b) hoax: falsehood deliberately fabricated to masquerade as the truth, e.g., fake scandalous reports, rumours, urban legends, and c) propaganda: refers to bias news designed to convince or persuade the reader to adopt a political or social agenda (Rashkin et al., 2017). Recently, some other researchers had included the category of ‘clickbait’. As stated by Vishwakarma and Jain (2020), (d) clickbaits include news articles that are completely false and are only created to attract the users’ attention upon clicking on the link, and therefore increase the internet traffic on these pages to add to the revenue. According to Volkova et al. (2017), the intent behind propaganda and clickbait varies from opinion manipulation and attention redirection to monetization and traffic attraction. It is also worth mentioning that from all the categories already described, satirical news and hoaxes might be the most harmful, especially when they are shared out of context (Conroy et al., 2015). The following section will briefly describe a range of datasets that have been previously offered as benchmarks.

2 Repository of Fake News Detection Datasets

Many scholars have intended to create datasets that comprise social media statements, such as those made by politicians or the public with information about their truthfulness. One of the earliest works on the automatic detection of fake news was provided by Vlachos and Riedel (2014). These authors were the first to produce a public fake news detection and fact-checking dataset, however, it only contained 221 statements extracted from PolitiFact and Channel. Subsequently, Mitra and Gilbert (2015) created CREDBANK[13], a large-scale dataset consisting of sixty million tweets related to over one thousand news events. Each event was assessed for credibility by thirty annotators from Amazon Mechanical Turk. Nonetheless, since CREDBANK dataset

[13] http://compsocial.github.io/CREDBANK-data/

was initially collected for assessing tweet credibility, this dataset has been proved to be ineffective for fake news detection.

Similarly, the LIAR[14] dataset is a well-known repository, which comprises over 12,000 human-labelled short statements with six fine-grained gradations of veracity produced from PolitiFact's database. Given the fact that this dataset only comprises brief sentences, using style-based techniques to detect deception is problematic, because it requires more content information. Another paramount dataset is BuzzFeedNews[15], which integrates 2,282 Facebook posts over a week before the 2016 US Presidential Elections. Every post was fact-checked by five BuzzFeed journalists: 826 mainstream, 356 left-wing, and 545 right-wing articles. The interesting aspect of this dataset is that it considers fake news from a more social media-oriented perspective. Nonetheless, it is important to highlight that it only contains headlines and text for each news piece and covers news articles from only nine news agencies. Similarly, the BuzzFace[16] dataset (Santia and Williams, 2018) is an extension of the BuzzFeedNews dataset, including comments related to news articles extracted from Facebook. The dataset comprises 2,263 news articles and 1.6 million comments related to news content.

Other existing datasets are BS Detector[17], FakeNewsNet[18], FEVER[19], PHEME[20], FA-KES and ISOT[21]. BS Detector dataset was collected through a browser extension named BS detector, which was developed for checking news veracity. It analyses the links on a webpage for references and checks them against a manually compiled list of domains. Hence, instead of using human expert annotators, BS Detector data is collected and annotated by developed news veracity checking tools. The FakeNewsNet repository (Shu et al., 2018), comprises nearly 24,000 news articles including information related to news content, social context, and spatiotemporal information from different news domains such as political and entertainment sources extracted from PolitiFact and GossipCop. The information content was annotated by journalists and domain experts. On the other hand, Thorne et al. (2018) created the FEVER (Fact Extraction and VERification) dataset. They included their own 188,445 produced statements by modifying Wikipedia sentences, and then providing evidence for or against each article. These articles are classified into three classes (Supported, Refuted or NotEnoughInfo). PHEME (Kochkina et al., 2018) dataset was created for rumour detection and veracity classification. It contains 330 twitter conversations on rumour tweets (297 in English, and 33 in German) which are based on linguistic set.

The FA-KES dataset (Salem et al., 2019) includes 804 news articles (of which 376 are fraudulent) related to the Syrian war. For the fact-checking labelling process, the

[14] https://paperswithcode.com/dataset/liar

[15] https://github.com/BuzzFeedNews

[16] https://dataverse.mpi-sws.org/dataverse/icwsm18

[17] https://gitlab.com/selfagencyllc/bs-detector/bs-detector

[18] https://www.kaggle.com/mdepak/fakenewsnet

[19] https://fever.ai/dataset/fever.html

[20] https://www.pheme.eu/2016/06/13/pheme-rumour-dataset-support-certainty-and-evidentiality/

[21] https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php

Syrian Violation Documentation Center (VDC) was used, with semi-supervised annotation. Despite being a small-scale dataset, it can be used for fake news detection for other related domains than war-related news. Finally, the ISOT Fake News dataset (Ahmed et al., 2017) integrates 45,000 news articles, equally distributed to the true and fake categories. Whereas the truthful articles were gathered from the Reuters website, the false information was obtained from different sites identified as fake by Wikipedia and Politifact. Both the FA-KES and ISOT datasets, comprise the whole content of each article, as well as the title, date, and topic. The primary article themes are politics and international news, and the dates range from 2016 to 2017. Table 1.1 provides a comparative summary of the previously described datasets.

Fake News Datasets					
Datasets	**Authors/ years**	**Content**	**Data size**	**Label**	**Annotation**
LIAR	Wang (2017)	Short statements	12,836	Six	Editors, journalists.
BUZZFEEDNEWS	2016	Facebook posts	2,282	Four	Journalists
BUZZFACE	Santia and Williams (2018)	Facebook posts	2,263	Four	Journalists
CREDBANK	Mitra and Gilbert (2015)	Tweets	60 million	Five	Crowdsourcing
BS DETECTOR	2016	Articles	Unknown	10 different types	Veracity checking tools
FAKENEWSNET	Shu et al. (2018)	Articles	23,921	Fake or Real	Journalists and domain experts
FEVER	Thorne et al. (2018)	Short statements (Wikipedia)	185,445	Three	Annotators
PHEME	Kochkina et al. (2018)	Tweets	330	True or False	Journalists
FA-KES	Salem et al. (2019)	News articles	804	True or Fake	Semi-supervised annotations
ISOT FAKE NEWS DATASET	Ahmed, Traore, Saad (2017)	News articles extracted from Reuters.com	45,000	Fake or Real	Journalists

Table 1.1: Comparison of existing fake news detection datasets

The quantity of research being done on automatic fake news detection is rising rapidly. In the past years, several approaches have been used to detect and address the problem of fake news detection. These approaches could be categorised as: 1) linguistic-based, or 2) network-based.

3 State of the art: An overview of different approaches

3.1 Linguistic-based approaches

From a linguistic perspective, deceptive messages are extracted and analysed according to specific language patterns associated with deception, such as the presence of emotive language, spelling mistakes or informal language. This approach could be broadly divided into three language levels based on textual features: (a) lexicon, (b) syntax, and (c) semantic. Typical lexical features include character and word-level attributes, such as the number and frequency of unique words in the text (i.e., number of demonstrative pronouns, first-person pronouns, punctuation marks count). These frequency statistics are commonly obtained by means of the Bag-Of-Word (BOW) models (Mohseni et al., 2019);(Zhou et al., 2019). On the other hand, the main aim at the syntax level is to assess sentence-level features, including bag-of-words approaches, “n-grams” and part-of-speech (POS) tagging, as can be seen in Zhou et al. (2019). Therefore, syntactic features or attributes may refer to the number of content words, syllables per sentence, as well as the tags of word categories (such as noun, verb, adjective). Regarding the semantic level, such frequencies may be attributed to lexicons or phrases that fall into each psycho-linguistic category, such as those defined in Linguistic Inquiry and Word Count (LIWC) (Pérez-Rosas et al., 2017), or into each self-defined psycho-linguistic attribute. An example would be sentiment analysis or opinion mining.

Investigating fake news detection on the basis of text has been explored by multiple academics. Reis et al. (2019) employed several classic and state-of-the-art classifiers, namely k-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forests (RF), Support Vector Machine (SVM), and XGBoost (XGB). Throughout the application of handcrafted features (subjectivity, lexical, syntactic, psycholinguistic and semantic), their results demonstrated that the prediction performance of proposed features combined with current classifiers exhibited a useful degree of discriminative ability for detecting fake news. Their most substantial classification results were obtained with RF (85%) and XGB (86%). However, it is important to mention that whilst the model could correctly detect nearly all fake news in their data, it tended to misclassify around 40% of legitimate news.

Benchmark research on fake news was conducted by Gravanis et al. (2019), who concluded that combining an enhanced linguistic feature set with powerful machine learning models was sufficient to effectively classify fake news. In their study, they introduced a new corpus, the "UNBiased" (UNB) dataset, which comprises several news sources and fulfils various standards and rules to prevent biased results in fake news classification tasks. According to the authors, the standards and rules that should be followed are: each fake news article must be annotated by experts, originate from several sources, be published by credible journalism organizations, and have a pluralistic collection with diverse categories. The proposed features combined with ML algorithms obtained accuracy up to 95% over five different datasets; the most accurate results were achieved with AdaBoost, followed by SVM and Bagging algorithms. Hence, their findings confirmed the efficacy of assessing news articles' truthfulness throughout the proper selection of specific features and ML algorithms. Similarly, Hossain et al. (2020), who presented the first labelled fake news dataset in Bangla, observed that lexical features performed better than other linguistic features; character-level features proved to be more significant than word-level features. Another key finding was that the use of punctuation marks in fake news was more frequent than in genuine news. Their evaluation of linear classifiers (SVM, LR, RF) and neural network-based models suggested that linear classifiers with traditional linguistic features performed better than neural network-based models. Their experiments demonstrated that SVM (91% F1-score) out-performed LR and RF by a significant margin.

On the other hand, Zhou and Zafarani (2020) investigated potential similarities in fake news and their apparent relationship to deception and clickbait. The authors presented a model based on the theory of detecting false news, in which a piece of news article is analysed at the lexical, syntactic, and semantic levels, in contrast to techniques that focus on news content (e.g., content quality, sentiment, quantity, and readability). The researchers evaluated SVM, RF, Gradient Boosting (XGB), LR and NB on the FakeNewsNet dataset. In their experiment, it was noticed that Word2Vec features performed better with linear SVM, whereas features based on Doc2Vec obtained more accurate results using XGB. Their findings proved that the suggested technique could outperform the state-of-the-art, in addition to detecting fake news early, even when the content is limited. Nonetheless, the experiment did not analyse possible linguistic structures divergences, and no language other than English was addressed.

With the major aim to identify similar universal characteristics that are independent of culture, Gruppi et al. (2018) compiled datasets in Portuguese and English. The authors discovered that the attributes of non-genuine articles followed a similar pattern in both languages, indicating the presence of related stylistic characteristics in both languages when separating legitimate and fake news. Thus confirming the existence of writing style differences between real and fake news. Nonetheless, their results were limited to only two languages, which is not sufficient to conclude that stylometric patterns might be observed in multiple languages. Furthermore, it is worth considering the researchers' reliability analysis of the articles, disregarding the verification or validity of the information and the analysis of other relevant features such as satire.

3.2 Neural network-based approaches

Contrary to the classical feature-based model, deep learning is advantageous since it does not require handcrafted features. Alternatively, in a network-based approach, language styles can be captured via word embeddings and recurrent networks. Recently, researchers have experimented with a variety of deep learning algorithms to detect fake news. Convolutional Neural Networks (CNN) and Long Short-Term Memory Units (LSTM) are two examples. Despite being intended for machine vision applications, CNNs tend to be very effective. Nonetheless, LSTMs occasionally outperform CNNs, since LSTM solves the vanishing gradient problem. These approaches determine the meaning of a word while considering its context; in contrast to traditional models, they perform effectively with vast amounts of text data. Attention mechanisms, in particular, have emerged as one of the most potent approaches in natural language processing. They are mostly used in conjunction with Recurrent Neural Networks (RNN) to forecast the most important information in an input sequence (Trueman et al., 2021).

Kaliyar et al. (2020), for example, built a deep learning model (FNDNet) to automatically learn the discriminatory features for fake news classification through a CNN with multiple hidden layers. FNDNet was analysed using GloVe as a pre-trained word embedding (100-dimensional version). By setting the dropout to 0.4 and using the ReLu (Rectified Linear Unit) function, their classification model obtained an accuracy of 98.36% on the test data. In their experiment, the authors used the Fake-News dataset, which consists of vast amounts of fake and real news articles related to the 2016 U.S. General Presidential Election. Successful results were also achieved by Nasir et al. (2021), who proposed a novel hybrid deep learning model on two fake news datasets (ISOT and FA-KES). Their model implements a CNN layer of Conv1D for processing the input vectors and extracting the local features at the text level (using GloVe pre-trained word embeddings). Subsequently, the RNN layer utilizes the extracted features and learns the long-term dependencies of the local features of news articles that classify them as fake or real. The activation function of choice was also ReLU. Obtaining an accuracy of 99% in the ISOT dataset, this hybrid model demonstrated that considerably better results can be obtained with hybrid baseline methods.

The benchmark research of Khan et al. (2019) evaluated some traditional models (NB, LR SVM), and neural network-based classifiers (LSTM, Bi-LSTM, Conv-LSTM) on three datasets: LIAR, Fake or Real News from Kaggle, and a Combined corpus with 80,000 samples. In their study, the authors extracted lexical and sentiment features, n-grams, Empath generated features, and pre-trained word embeddings (100 dimensions GloVe). Their findings showed that even though neural networks perform better on larger datasets (e.g., 95% accuracy with BLSTM and Conv-LSTM), NB may outperform neural networks on smaller datasets with up to 95% accuracy. Another significant study in this field was conducted by, Kaliyar et al. (2021), who implemented a deep learning approach by combining BERT (Bidirectional Encoder Representations from Transformers) and three parallel blocks of 1d-CNN, having different kernel sizes and filters with the BERT. Their model, FakeBERT was trained on the Fake-News dataset.

The findings demonstrated that FakeBERT outperformed current state-of-the-art models with an accuracy of 98.90%. On the other hand,

Overall, three concluding ideas could be emphasized from this literature survey. Firstly, supervised classification employing features of varying complexity is a common approach to the problem of fake news identification. However, supervised approaches require vast amounts of labelled data. Secondly, since most datasets are not large enough to train complex neural networks, classical algorithms are frequently employed to tackle the issue of fake news detection. Finally, it is also noteworthy to mention that although annotated data is far more widely available in English, it is practically non-existent in the great majority of languages.

4 Spanish and French Fake News Detection

4.1 Fake News Detection in Spanish

While there are existing tools for the detection of fake news in Spanish, the majority of them perform the verification process manually; via extensive investigations by the researchers and users. This suggests that the identification of deceptive content is subjective and time-consuming. Examples of platforms that aims to tackle the fake news problems throughout journalists' manual validation are: VerificadoMX3[22] with a journalist approach, Maldito Bulo[23] to refute rumours and fake news on social media, and Caza Hoax[24], which is a Hispanic-based community specialized in unmasking misleading information on the internet.

Despite all the efforts and improvements, there is still not sufficient, publicly available, or adequately labelled Spanish corpus to train ML or DL models to effectively detect fake news articles. The benchmark research of Posadas-Durán et al. (2019), is the only officially recognized study of fake news detection in Spanish. The authors developed and made publicly available the first corpus written in Spanish, named Spanish Fake News Corpus[25]. This corpus compiles news from a variety of online sources, including websites of established newspapers, media companies, specifically dedicated for fake news verification, and websites identified by different journalists as sites that routinely produce fake news. The corpus contains 971 samples labelled as True or Fake. In their experiment, the researchers implemented four machine learning classifiers: SVM with linear kernel, LR, RF and boosting (BO). These classification algorithms were trained on three lexical feature representations: standard bag-of-words model, POS tags, and n-grams (with n varying from 3 to 5, and n-grams combination). The researchers assessed the performance of character and word n-grams type when incorporating and omitting stop words. The most accurate results on the test set were obtained with character 4-grams without removing the stop words with the Boosting algorithm

[22] https://verificado.mx/

[23] https://maldita.es/malditobulo/1

[24] cazahoax.com

[25] https://github. com/jpposadas/FakeNewsCorpusSpanish

(77.28%). In their experiment, it was discovered that the models based on character n-grams achieved better results, and that the exclusion of stop words decreased the performance of the classifiers. Therefore, their classification results proved the corpus potential for building future fake news detection models, as well as the possibility to achieve very high accuracy using classic state-of-the-art classifiers with lexical features.

A different approach was conducted by Huang et al. (2021), who also used the Spanish Fake News Corpus[26] (Posadas-Durán et al., 2019) to train a BERT model to detect Spanish fake news. Specifically, the authors implemented BERT for obtaining two text embeddings (which were called Head Embedding and Tail Embedding). They also added a Sample Memory with an attention mechanism to capture richer semantic information in long news texts. The Sample Memory consisted of a matrix parameter initialized by sample representation, and combined with the attention mechanism. The researchers calculated the dot-product attention between the result and Sample Memory utils to obtain Memory Embedding. Finally, the Beginning Embedding, End Embedding, and Memory Embedding were stitched together to obtain the output result. Their model proved successful in capturing the relationship information between samples which strengthens the model's robustness in the inference stage. Their project's findings achieved first place in IberLEF 2021 with 86.44% accuracy.

4.2 Fake News Detection in French

Given the current limitations in Fake News Detection research, and the complexity to detect deliberate fake news, French-related research tends to be focused on the phenomenon of satirical fake news. To the researcher's knowledge, the most illustrious experiment on this domain was conducted by Liu et al. (2019). The authors created their own dataset comprising 5,682 samples extracted from six different sources that included the articles' full text, topic and labels indicating their class (real or satire). They utilized Term Frequency-Inverse Document Frequency (TF-IDF) method to compute term-frequency and inverse document-frequency with n-grams (1,5), and ended with a feature vector length of 3,000. From the models implemented, the Logistic Regression algorithm performed best, with an accuracy of 92.17%, followed by Neural Networks (94.68%), and SVM (93.58%).

A different approach was addressed by Guibon et al. (2019), who employed a dataset composed of French and English samples with three possible classes: Fake News, Trusted News and Satire contents. The corpus selected for this study was the vaccination fake news dataset provided by the Storyzy company (an independent adtech company specialized in quote extraction to detect Fake News in real time with NLP). Natural Language Toolkit(NLTK) FrenchStemmer was used to obtain the stem of French words and therefore reduce the vocabulary size. It should be noted that their data representation remains language dependent, since it relies on specific language pre-training methods such as word vectors (Word2Vec, FastText) or term frequencies

[26] https://github. com/jpposadas/FakeNewsCorpusSpanish

(TF-IDF). Their findings proved that CNN tends to work more effectively for discrimination of the larger classes, with 95.78% accuracy in trusted news, and 92.71% in fakes news. On the other hand, the gradient boosting decision tree with a feature stacking approach generated more accurate results for satire detection (93.75%).

5 Conclusion

This literature review leads to a wide range of conclusions. Firstly, it is relevant to mention that considerable efforts have been made to automatically detect all possible types of fake news, such as hoax, propaganda, satire and clickbait. However, in any supervised approach, annotated corpora are required; which poses an additional challenge given the scarcity of relevant labelled fake news databases, and the fact that most available resources are in the English language. Because fake news is prevalent in all languages, it is critical to effectively address this issue from different perspectives. It is also worth mentioning how every considered dataset exhibits some limitations. The LIAR dataset, for example, comprises very concise statements, whereas the BuzzFeed-News dataset only contains headlines and texts.

Recent studies demonstrate that fake news detection tasks rely mainly on linguistic features as well as semantic patterns at the word level. Extensive academic research has manifested how a combination of enhanced linguistic features together with powerful machine learning models is capable of reaching high accuracy. At the textual level, for example, lexical features have proved to give the most remarkable results (Hossain et al., 2020); (Posadas-Durán et al., 2019). On the other hand, several studies (Liu et al., 2019),(Guibon et al., 2019) have shown that great results can be obtained by converting tokenized texts into features with TF-IDF. The most frequent linguistic-based classifiers for fake news detection are Support Vector Machine, Logistic Regression and Random Forest. Regarding deep learning models, RNN and CNN have been frequently used with pre-trained word embeddings (mostly GloVe and Word2Vec) to solve this classification problem. Although neural network-based models achieve considerable results, sometimes they might occasionally be outperformed by linguistic-based approaches, such as in Liu et al. (2019), where a LR classifier recorded higher accuracy than neural networks models.

Alternatively, it is worth mentioning that the conducted research in Spanish and French is practically non-existent. In relation to Spanish fake news research, there is only one recognized study conducted by Posadas-Durán et al. (2019), who also developed and made publicly available the first corpus written in Spanish, named Spanish Fake News Corpus. However, they only implemented linguistic-based models. Using the Spanish Fake News Corpus, Huang et al. (2021) implemented a BERT model which outperformed with 86% accuracy the results achieved by Posadas-Durán et al. (2019). Similarly, French research on the topic is still limited and mainly focused on the phenomenon of satirical French news on social media. On the other hand, it is important to highlight that there have been some attempts for cross-lingual and multilingual fake news detection. Guibon et al. (2019), for example, addressed a dataset composed of French and English fake news related to vaccination, and obtained 95.78% accuracy using CNN. This dataset has been reported to be one of the first attempts to create a more generic fake news detection approach that is not limited to any specific language. Table 1.2 includes a description of the datasets, and the accuracy reported by the different models presented in this literature review.

Fake News Datasets							
Language	**Authors/ years**	**Datasets**	**Data Type**	**Data size**	**Classification**	**Features**	**Models and accuracy**
English	Reis et al. (2019)	BuzzFace	Facebook posts	2,263	Four: Mostly true, mostly false, mixture of true and false, non-factual	Syntactic, lexical, semantic and psycholinguistic features.	K-NN - 80% NB - 72% RF - 85% SVM - 79% XGB - 86%.
English	Gravanis et al (2019)	UNBiased (UNB) dataset	Kaggle-EXT, McIntire, BuzzFeed, Politifact, UNB.	3,004	Binary: Real or False	Linguistic features enhanced with word embeddings	AdaBoost - 95% SVM - 95% Bagging - 94% K-NN - 92% NB - 88% DT - 85%
Bangla	Hossain et al. (2020)	BanFakeNews	Web articles	50,000	Binary: Authentic or Fake	Lexical, semantic, syntactic, meta-data, word embeddings	F1-score: SVM - 91% LR, RF CNN - 59% LSTM - 53% BERT - 68%
English	Zhou et al. (2020)	FakeNewsNet dataset	PolitiFact and GossipCop: political news	23,921	Binary: True or Fake	Lexical, syntactic, semantic and latent features (i.e., XGBoost)	RF - 87% XGB - 89%
Brazilian Portuguese and American English	Gruppi, Horne, and Adali (2018)	NELA2017 dataset (US). Media sources (BR).	Political articles, media sources	5,511 (BR) 2,841 (US)	Reliable, Unreliable, Satire	Complexity, style, linguistic, and psychological	SVM-85% (BR) SVM-72% (US)
English	Kaliyar et al. (2020)	FAKE-NEWS	Political articles	20,761	Binary: True or False	Pre-trained word embeddings	CNN with multiple hidden layers - 98.36%

Language	Authors years	Datasets	Data Type	Data size	Classification	Features	Models and accuracy
English	Nasir et al. (2021)	ISOT and FA-KES	News articles	ISOT: 45,000 FA-KES: 804	Binary: True or Fake	Pre-trained word embeddings (GloVe)	Hybrid CNN-RNN model - 99% (ISOT)
English	Kaliyar (2021)	FAKE-NEWS	Political articles	20,761	Binary: True or False	Word embeddings	BERT and 1d-CNN - 98.90%
English	Khan et al. (2019)	LIAR, Fake or Real News (Kaggle), and a Combined corpus	Short statements, political articles	LIAR=12,800 Fake or Real=7,800 Combined corpus=80,000	Fake or Real: True or Fake LIAR: Six classes	Lexical and sentiment features, n-grams, Empath generated features, and pre-trained word embeddings	Bi-LSTM and C-LSTM - 95% NB - 95%
English	Elhadad et al. (2019)	ISOT, FA-KES and LIAR	News articles and short statements	ISOT=45,000 FA-KES=804 LIAR=12,800	FA-KES: True or Fake	Content features and textual metadata	ISOT: 100% LIAR: 62% FA-KES: 58%
Spanish	Posadas-Durán et al (2019)	FakeNewsCorpus Spanish	News websites	971	Binary: True or Fake	Lexical fetures (character n-grams, POS n-grams, including and excluding stop words)	SVM - 71.52% LR - 73.55% RF - 76.94% BO - 72.54%
Spanish	Huang et al. (2021)	FakeNewsCorpus Spanish	News websites	971	Binary: True or Fake	Text embeddings	BERT - 86.44%
French	Liu et al. (2019)	Twitter posts	French satirical stories	5,682	Binary	Word vectors, TF-IDF with n-grams (1,5)	NNs- 94.68% SVM-93.58% RF- 95.18% LR- 95.25% NB- 92.32%
French and English	Guibon et al. (2019)	Vaccination fake news dataset provided by the Storyzy company	Web articles and French transcripts from YouTube	5,105 (EN) 1,253 (FR)	Fake News, Trusted and Satire	Word2Vec, FastText, TF-IDF, word embeddings and lexical features containing redundancy	CNN- 95.78% XGB- 93.75% (Satire)

Table 1.2: Summary of state-of-the-art research conducted on Fake News.

Chapter 2

Methodology

Due to the scarcity of research on fake news detection in languages other than English, one of the primary contributions of this research project is to provide a comparison of automatic fake news detection in Spanish and French. Therefore, the subsequent sections will explain the various processes involved in developing a classifier for fake news detection in the two languages mentioned previously. The first two sections provide a description of the selected datasets, and the pre-processing steps applied to clean the data for further analysis. The third section introduces a linguistic-based model proposed for fake classification. Subsequently, the fourth section presents a linguistic-statistical analysis of the two datasets; elaborated in an attempt to identify the most significant linguistic features of Spanish and French fake news as compared to real news. Conclusively, the final section includes a comparison of deep learning classification models with the linguistic-based models presented in the previous section. The task is defined as a binary classification problem (True: '0' or Fake: '1').

The machine and deep learning models implemented in this thesis have been developed on Google Colab Pro (Python 3.6.9) using GPU environment. Among the Python libraries employed on this project, Pandas[1] package was used for reading the datasets and processing arrays; Numpy[2] library was also used for processing arrays. Regarding the cleaning and pre-processing steps, re[3], spaCy[4] and NLTK libraries were utilized. Scikit-learn[5] (sklearn) package was applied to obtain the results and vectorization in the baseline classifiers. On the other hand, Keras[6] and Tensorflow (already integrated into Keras) frameworks were used for the implementation of the neural network models. As for the data visualization stage, wordcloud (python library for generating wordclouds), matplotlib (plotting library for Python), and scikitplot were utilized to construct different plots, tables, graphs, and images.

[1] https://pandas.pydata.org/

[2] https://numpy.org/

[3] https://docs.python.org/3/library/re.html

[4] https://spacy.io/

[5] https://scikit-learn.org/stable/index.html

[6] https://keras.io/

1 Data Description

One of the major challenges for the task of fake news detection is to compile reliable news data annotated as fake; hence, four existing corpora were selected. For Spanish, FakeNewsCorpusSpanish (Posadas-Durán et al., 2019) and a Spanish fake news dataset obtained from Kaggle[7], were chosen. As described in the previous chapter (subsection 4.1), The FakeNewsCorpusSpanish, by Posadas-Durán et al. (2019), contains 971 news divided into 491 Real news and 480 Fake news. The authors collected the samples from over one hundred different sources, providing the category of the news and applying a normalization process to their corpus, where specific elements were masked by common identifiers: 'NUMBER', 'EMAIL' 'URLs', 'PHONE', 'DOL' (dollar) and 'EUR' (euro) symbols. Furthermore, the researchers presented a detailed analysis of vocabulary overlap among categories. The overlap was calculated by dividing the intersection of both true and fake vocabulary between the joint vocabulary. The results demonstrated that the general vocabulary overlap between real and fake news was 27.68%. Since this corpus was not sufficient to train a neural network model, the researcher opted to increase its size by adding a second dataset. The second chosen Spanish corpus was created in 2018 and downloaded from Kaggle. It contains 993 samples labelled as 'True', and 956 as 'False'. One of the main drawbacks of using the latter corpus is that it consists of a collection of short samples that only includes the text and the category to which the articles belong. Hence, other necessary attributes, such as the source or topic should have been included to guarantee the veracity in the compilation process. Therefore, the only two current existing Spanish datasets for fake news detection were merged and used as our chosen corpus.

On the other hand, since current state-of-the-art French corpora for fake news detection is focused on sarcasm, the datasets selected for deliberate French fake news were obtained from GitHub users: FakeNewsDetectionFr[8] and FrenchFakeNews[9] The reason behind merging two French dataset was to avoid any specific writing style particular of any of the dataset, include more topic variety, and include different subcategories of fake news (e.g., disinformation, satire, clickbait). The first French fake news dataset is a collection of 2,423 web articles: 1,361 labelled as 'True' (0) and 1,071 labelled as 'Fake' (1). The genuine articles of the FakeNewsDetectionFr dataset were selected from five different French sources: Futura Sciences, Liberation, Le Figaro, 20Minutes and Le Monde. Similarly, the fake samples were collected from three different French Fake News websites: Le Gorafi, Nordpresse and Buzzbeed. The vast majority of these articles were published between 2019 and 2020, and are related to twenty-one different topics, among which science, technology, sport, politics, and economics could be highlighted. On the other hand, the FrenchFakeNews dataset comprises 26,834 articles, 13,416 belonging to the fake news category, and 11,968 to real news. Hence, by combining both corpora in each language, the final Spanish dataset contained 2,971 articles, whereas the final French dataset comprised 27,816 articles. Table 2.1 summarises the datasets

[7] https://www.kaggle.com/arseniitretiakov/noticias-falsas-en-espaol?select=fakes1000.csv

[8] https://github.com/jeugregg/FakeNewsDetectionFr

[9] https://github.com/akachbat/FrenchFakeNews

already described. The final corpora contained different types of fake news, such as satire, hoax, and propaganda. However, this project does not establish a distinction among different categories, but attempts to detect deliberate fake news in general. Table 2.1 provides a summary of the datasets selected for this research project together with the mean of words per article in each dataset.

Language	**Corpus**	**Size**	**Words Mean**	**Content**
Spanish	FakeNews CorpusSpanish	971 **Fake:** 480 **True:** 491	440	articles
Spanish	Kaggle	1,958 **Fake:** 965 **True:** 993	43	web articles
French	GitHub FakeNewsDetectionFr	2,432 **Fake:** 1,071 **True:** 1,361	566	news and parody articles
French	GitHub FrenchFakeNews	25,384 **Fake:** 13,416 **True:** 11,968	445	news articles

Table 2.1: Description of the datasets employed in this task

2 Data cleaning and pre-processing

The performance of any text classification model is highly dependent on the number of words in each corpus and the features created from those words. Content words (otherwise known as stop words) and other "noisy" elements increase feature dimensionality but do not usually aid in document differentiation. For this purpose, the researcher converted the Spanish datasets into ISO-8859-1 encoding format, and used the NLTK package in Python to clean the entire corpora. The cleaning of the datasets included converting words to lowercase and the deletion of non-informative textual features such as stop words (stopwords.words ('spanish'), stopwords.words('french')) and punctuation marks. Two versions of the same dataset were performed: with and without accents and stop words. Subsequently, all the tokens were lemmatized. Lemmatization implies converting each word to its root form, and therefore grouping different derived or inflected words into a single representation. Spacy lemmatizer was applied to the French (fr_core_news_sm) and Spanish (es_core_news _sm) datasets. Conclusively, the substitute function in regular expressions ('re.sub') was employed to omit multiple spaces ('\s*'), non-alphabetic Spanish ([\a-zá-ú]+) and French ([\a-zà-ùä-üâ-ûë-ü]+) characters, URLs, HTML tags, and the normalized expressions in the Spanish Fake News Corpus (i.e., 'NUMBER', 'URLs', 'PHONE'). Each of these textual features was replaced by a space in the two corpora. Accented vowels were converted to their non-accented respective form in Spanish (i.e., r 'á', 'a', text) and French (i.e., r 'ô', 'o', text).

The final phase consisted in assigning weights to words. When converting the news articles to a matrix of token counts, the researcher noticed that some French articles still contained some contractions of stop words, such as ‘l’ (le - the), and ‘d’ (de - of). Consequently, a final cleaning step was introduced to remove French contractions or additional words which do not provide any significant information. Moreover, these steps assisted in reducing the vocabulary size of the corpora, excluding any of the datasets’ specific writing styles; therefore, the classifier would not automatically assign labels to an article whenever such a feature exists. Following the implementation of the cleaning and pre-processing steps, the French corpus comprised 5,956,087 words, a mean of 214 words per article; whereas the Spanish dataset only contained 239,951 words, with a mean of 80 words per article.

3 Linguistic-based models

Once the articles had been pre-processed, the texts were tokenized and converted into features using two baseline models of the Sklearn library: Count Vectorizer (CV) and TF-IDF. Count Vectorizer converts a given set of strings into a matrix of token counts. The CV parameters were chosen based on the premise that words that appear in almost every document are unsuitable for classification because they do not provide significant information. Therefore, the max_features parameter was set to 5,000 in order to build a vocabulary that only considers the most frequent words as features. Similarly, the max_df feature value was set to 0.8 to ignore terms that have a document frequency strictly higher than the given threshold (corpus-specific stop words). The CV’s vector features were verified to ensure that the features in both datasets were indeed word names and not noise. Simultaneously, TF-IDF was implemented to convert both datasets to a matrix. TF-IDF is based on the frequency of a word in the corpus, and provides a numerical representation of how important a word is for statistical analysis. Aiming to reduce the input dimensions, the Logarithmic factor in TD-IDF mathematically penalizes the words that are too frequent or too rare in the corpus by giving them low TF-IDF scores. The max_features parameter was also set to 5,000.

The fit-transform function transformed the datasets’ words into frequency representations with CV, and into weighted frequency with TF-IDF values. The TF-IDF classifier was run on words, as well as on character and word n-gram vectors of a (2,3) range. The reason behind including n-grams, is due to how they add context prior to feature conversion. Accordingly, n-grams combine nearby words into single features, which gives context to words that may have limited meaning on their own. The train-test_split utility from the sklearn.model_selection library was used to randomly separate the samples of each dataset into 70% training set, 20% test set and 10% validation set. In order to solve this binary classification task, a Logistic Regression classifier was run on tokenized and segmented versions of the fake and real datasets. After obtaining a considerable high accuracy with the first baseline model, the researcher attempted to add various linguistic features, characteristically recognized in fake news articles, to the classification task. The linguistic features extracted are: slang, highly emotional words, character and word count, title count, uppercase count, punctuation count, and word density.

3.1 Semantic features

Informality

By observing the samples in both corpora, it was found that some fake articles contained informal language patterns such as slang and profanity words. Consequently, this feature was added to the models. According to Burfoot and Baldwin (2009) informal language is much more common in fake articles, which implies that true news articles tend to avoid slang, profanity, and exaggerated language. The researcher measured the informality of each article following the equation applied by Burfoot and Baldwin (2009), where ‘T’ refers to the set of tokens in the article and ‘s’ is a function taking the value ‘1’ if the token appears in the slang dictionary, and ‘0’ if it does not appear. See equation 2.1.

$$i^{def} = \frac{1}{|T|} \sum_{t \in (T)} s(t) \tag{2.1}$$

Therefore, for this purpose, the researcher compiled two lists of slang and profanity words and phrases; 903 in Spanish and 1,927 in French. The list of Spanish words was compiled from two main websites: Language Realm[10] (Spaniard), and baselang[11] (Mexican Spanish). Similarly, the list of French words and expressions was created from Language Realm[12]. Correspondingly, the informality of each article was measured according to the equation previously described. To accurately identify the words and expressions present in the list, a different cleaning process was applied to the French and Spanish corpora, where words were lemmatized, but stop words and accents were not removed, with the aim of not altering the meaning of certain expressions, such as ‘a punto de’ (on the point of).

Emotion analysis:

Genuine articles tend to express opinions objectively to avoid implied bias. On the contrary, fake news may include emotional language for entertainment. As stated by Bhutani et al. (2019) fake news categorization may depend on whether the writer’s attitude towards a particular topic is positive, negative, or neutral. However, when an individual or entity makes a false statement or poses allegations against someone else, such as a political party, the sentiment of the statement tends to be negative, which indicates that the news is most probably fake. In their investigation, Bhutani et al. (2019) incorporated sentiment as an important feature to improve their accuracy in

[10] http://www.languagerealm.com/spanish/spanishslang$_z.php$

[11] https://baselang.com/blog/vocabulary/mexican-slang/

[12] http://www.languagerealm.com/french/frenchslang.php

detecting fake news. Another example of existing works that suggested the potential of sentiment features for fake news detection is Khan et al. (2019), who attempted to build a model for fake news detection, where the sentiment of every article was classified as positive, negative or neutral. To extract sentiment features, they used SentimentIntensityAnalyzer function of python NLTK library.

Emotion analysis refers to recent work development in Sentiment Analysis of text used in NLP. Although it is not completely accurate to state that fake news articles' content is chiefly negative, this feature is claimed to be paramount in determining whether the news article is fake or not. Thence, it is hypothesized that the sentiment reflected in writing a news article could serve as a pivotal deciding factor in the procedure of characterizing the news into fake or real. To obtain the highly emotional negative and positive words in the Spanish dataset, the Spanish Emotion Lexicon[13] (SEL) resource (Sidorov et al., 2012) was employed. The SEL dataset comprises 2,036 words associated with specific emotion weights (0 to 1) and categories: Anger, Fear, Sadness, Joy, Surprise, and Disgust. Table 2.2 represents some samples included in the SEL corpus with their associated emotion category and measure of Probability Factor of Affective (PFA) use.

Word	**Anger**	**Fear**	**Sadness**	**Joy**	**Surprise**	**Disgust**
Felizmente	0	0	0	0.966	0	0
Sufrimiento	0	0	0.898	0	0	0
Temible	0	0.932	0	0	0	0
Resentimiento	0.731	0	0	0	0	0
Atónito	0	0	0	0	0.799	0
Nefasto	0	0	0	0	0	0.831

Table 2.2: Example of SEL words with the associated emotion and PFA

Instead of accumulating the weights to calculate the prevalence of each emotion within the text, the researcher opted to group the six emotions into a binary category: 'positive' or 'negative'. Therefore, each word's emotional value becomes binary. That is to say, the presence in the article of a word listed in the SEL dataset will be represented as '1'. The total number of emotional words are counted to obtain an Emotion Score that will then be normalized by dividing it by each article's total number of words. Similarly, the French Expanded Emotion Lexicon (FEEL) polarity dictionary, by Abdaoui et al. (2017), was also employed to calculate the normalized frequency of positive and negative emotional words in the French corpus. The FEEL dataset contains 14,127 words and expressions labelled as 'positive' or 'negative'.

[13] https://mailman.uib.no/public/corpora/2012-December/016707.html

3.2 Lexical features

Previous work on fake news detection (Rubin et al., 2016) suggests that the use of punctuation might be useful to differentiate deceptive from truthful texts. Accordingly, the number of punctuation marks per article was calculated. Based on previous studies (Choudhary and Arora, 2021), the researcher also extracted various lexical-based features: words and characters count, title and uppercase word count, as well as word density (average word length). According to Choudhary and Arora (2021), language features comprise several linguistic dimensions which follow a particular pattern to classify fake news. This stage consists of the extraction of paramount count statistical evidence. During the creation of fake news, the author intentionally employs title words, uppercase words, and even considers their content length and corresponding word density. The distribution of the mean of these features was included as new features to the datasets. Table 2.3 includes the definition of the already mentioned lexical features. The researcher alternatively adopted a state-of-the-art classification of news articles via a deep learning approach with word embeddings that can capture both the lexical and semantic characteristics of the positive dataset. The deep learning models applied in the research are detailed in the subsequent sections.

Lexical-based features	**Definition**
Char count	Total nº. of characters with spaces per article
Word count	Total nº. of words in a given article
Uppercase word count	Total nº. of uppercase words per article
Word density	Character count divided by word count
Punctuation count	Total nº of punctuation marks per article

Table 2.3: Description of the extracted syntax-based features

4 Deep Learning models

Most of the applied features in traditional machine learning techniques must be identified by domain experts in order to reduce the complexity of the data and make patterns more obvious for learning algorithms to work. Conversely, Deep Learning models benefit from how they attempt to learn high-level features from data in an incremental manner, reducing the requirement for domain expertise, and without having the need to extract complex features. Regarding the deep learning approach, three neural network-based models were used, where the researcher strictly tuned all the hyperparameters of the classification models on the training dataset. The models were implemented using Keras, an open-source neural network Python library designed to provide fast experimentation with deep neural networks. Keras focuses on being modular, user-friendly, and extensible. The models were built using the Sequential architecture of the Keras library, which is centered on the creation of models as a linear stack of layers. The following sections details three neural architectures used for classification:

LSTM, BLSTM (Bidirectional Long Short-Term Memory) and CNN+BLSTM.

4.1 LSTM with CV

The first Sequential model was used with an input layer of vectors obtained with Count Vectorizer (max_features=3,000). A Long short-term memory (LSTM) network was implemented in this model. According to Greff et al. (2016), 'LSTMs have emerged as an effective and scalable model for several learning problems related to sequential data' (p.1). LSTM networks are a type of RNN that uses special units in addition to standard units. The reason behind the utilization of LSTM is due to how it manages to vanish and explode the gradient problem throughout the introduction of new internal mechanisms called gates, such as input and forget gates, which function to regulate how much of the information is added and how much is forgotten. From this information, it passes the relevant data down the long chain of sequences to make predictions. Therefore, the LSTM model takes the count vectorized sequence of words as the input and predicts if the article is real or fake.

To calculate the values for each output node ('0': True or '1': Fake) in the network, each input node (vector) is multiplied by the total sum of the weight and bias. The input layer is directly connected to the LSTM layer (100 cells). The LSTM layer passes the information to three Dense layers. In the Dense layer, all the neurons from the previous layers are densely connected. Recent research (Tan et al., 2019), has suggested the implementation of one or two dense layers to prevent over-fitting in neural network models. Therefore, the researcher included three dense layers with a diverse number of filters: 64, 32 and 8 neurons, respectively. Rectified linear unit was chosen as the activation function for the three layers. The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons simultaneously. Finally, the last Dense layer is connected to the output layer with a sigmoid function. The model was computed with ADAM optimizer, and binary cross-entropy loss function. ADAM optimizer with learning rate 0.0001 was applied to minimize categorical cross entropy loss in the Spanish dataset. This model was trained for 15 and 10 epochs with the Spanish and French datasets, respectively.

4.2 BLSTM with word embeddings

The second proposed model consists of a Bidirectional Long Short-Term Memory. BLSTM outperforms both unidirectional LSTM and conventional Recurrent Neural Networks (Graves et al., 2005). The architecture's input layer contains lexical unigrams, in which each word is represented as a vector using a word-embedding approach. Firstly, the pre-processed articles are vectorized into lists of integers using the Keras library's Tokenizer function. Each integer maps to a value in a dictionary that encodes the complete corpus. In this dictionary, the keys refer to the vocabulary items themselves, and the indexing is ordered after the most common words in each article. This task is repeated independently for training and test texts in the two datasets. The maximum sequences length found were 1,391 and 3,521 for the Spanish and French datasets, respectively. Because the length of each sequence varies, post-padding was applied in order to convert all the sequences to the same length by appending zero values at the end of each sequence that is shorter than the fixed length.

The model comprises multiple layers. The first layer of the neural network is the Embedding layer created from the datasets' vocabulary (using a window of size 2), where each word is represented by 300 vectors. The next layer is the BLSTM, with 100 neurons and ReLu function. Subsequently, the BLSTM layer is connected to a Dense or Fully-connected layer with 32 cells. As a technique to avoid overfitting, the researcher opted to add a Dropout(0.3) layer between the LSTM and Dense layers. Finally, the trained feature vectors are classified using a final Dense layer, which reduces the dimension of the output space to 1, which refers to the classification label (Fake or not Fake). This model employed the same activation and loss functions as in the previous LSTM model. The training was performed for 50 epochs in the Spanish dataset with a batch size of 32, and for 7 epochs and a batch size of 128 in the French dataset.

4.3 A hybrid CNN+BLSTM with pre-trained word embeddings

The third model consists of a hybrid architecture that employs the Convolutional Neural Network's ability to extract local features and the BLSTMs networks to learn bidirectional long-term dependencies. This model uses pre-trained word embeddings of 300 dimensions. The FastText embeddings from SUC[14] was used for the Spanish repository. This dataset (3.4 GB) was created by José Cañete at BotCenter using the Spanish Unannotated Corpora, which contains 3 billion words. The word embeddings were computed with 300 dimensions and 1,313,423 vectors. In contrast, the French pre-trained word embeddings were obtained from FastText[15] (4.2 GB). The French word embeddings were created on Common Crawl and Wikipedia using FastText. These models were trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives.

This Sequential model is composed by 6 layers of neurons. The first layer of this hybrid neural network is Keras' embedding layer. This is the input layer where pre-trained word embeddings are used by providing the generated embedding matrix. In order to process the input vectors and extract the local features that reside at the textual level, a one-dimensional Convolutional Neural Network (Conv1D) layer is implemented after the Embedding Layer (with 64 filters of size 2 to extract local features). The Rectified Linear Unit activation function is employed by default. A Maxpooling1D Layer is also included, with the final aim to down-sample the input representation by taking the maximum value over a spatial window of size 4 (pool_size= 4). The pool feature maps are fed into the following two BLSTM layers of 40 units, which outputs the

[14] https://github.com/dccuchile/spanish-word-embeddings

[15] https://fasttext.cc/docs/en/crawl-vectors.html

bidirectional long-term dependent features of the input feature maps, while retaining a memory to identify news articles as false or real. These layers also apply the ReLu activation function.

Subsequently, the neural network integrates two Dense layers of 32 and 16 cells, respectively, with the ReLu activation function. Finally, the trained feature vectors are classified using a Dense layer, whose dimensions are reduced to 1. Three Dropout (0.1) layers were included between the different layers also to prevent overfitting. The Sigmoid activation function is used in this layer. Following the same previous process, the model was trained using Adam optimizer (learning rate of 0.0001 in the Spanish corpus) and binary cross-entropy as the loss function. The training was performed for 10 epochs using a batch size of 128 with the French dataset, and for 25 epochs and a batch size of 32 with the Spanish dataset. A comparison of the different results obtained with the neural architectures and the Logistic Regression baseline models is presented in the following chapter.

Chapter 3

Results and Findings

1 Baseline models

By examining the results obtained with the Spanish dataset, without considering the hand-crafted features, it was observed that word-level TF-IDF achieved the highest accuracy of 76.8% in the validation set, whereas the CV recorded 69% accuracy. Notwithstanding, the LR classifier performed more accurately in the French dataset, where the most accurate results were achieved with the Count Vectorizer: nearly 95%, followed by 94.8% with the TF-IDF model at the word level. Figure 3.1 illustrates two confusion matrices of the TF-IDF and CV models' performance on the validation set in the Spanish, and French datasets, respectively. These confusion matrices demonstrate that the classifiers' prominent errors with both models are Type II errors or False Negatives (FN), since it assigns real labels to fake articles. Whereas the TF-IDF model misclassified 28 False Negative (FN) and 27 False Positive (FP) Spanish articles, the CV misclassified 64 FN and 47 FP French pieces of news.

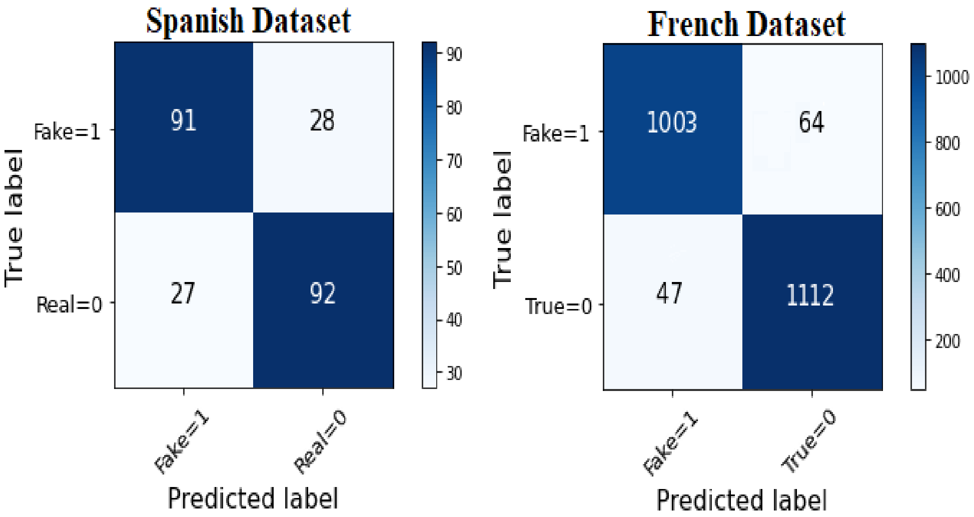


Figure 3.1: Confusion matrix of the TF-IDF model (Spanish dataset) and CV (French dataset)

Having added the informality and emotion features to the model, the researcher discovered that the model's accuracy remained the same in the Spanish dataset (75%). Correspondingly, the inclusion of these two additional attributes in the French dataset did not imply any significant difference in the performance of the models; the CV model still recorded 95% accuracy, although the number of FN instances was reduced by 1. The researcher observed over 50% of the Spanish articles rendered a 0 value in the slang feature; while nearly 25% of the French articles did not contain any informal words. Table 3.1 shows the slang results obtained with different measures in the two corpora. As it can be seen, the mean of slang words represents 0.03% and 0.003% in the French and Spanish datasets respectively. On the other hand, the maximum mean of slang words in an article is 0.48% in the French Dataset and 0.19% in the Spanish corpus.

SLANG	**French Dataset**	**Spanish Dataset**
count	27816.000000	2971.000000
mean	0.032545	0.003316
Std.	0.018639	0.013036
Min	0.000000	0.000000
25%	0.021583	0.000000
50%	0.031128	0.000000
75%	0.042555	0.010000
Max	0.488584	0.192308

Table 3.1: Calculation of the slang feature in the French and Spanish datasets

Regarding the inclusion of emotional words and expressions, the researcher noticed that 658 (out of 2,971) articles did not contain any of the words included in the lists. On the other hand, nearly 3,700 French articles (out of 27,816) did not include any emotional words. By computing the mean of highly emotional words per article, it was noted that, contrary to the researcher's expected results, positive and negative emotional words occurred more frequently in Real News in both datasets. Figure 3.2 represents the mean of the positive, negative, and total number of emotional words in Real and Fake articles in both datasets. The mean of emotional words per article in the Spanish dataset is 4.8 in Fake news, and 5.7 in Real news, while in the French dataset, it accounts for 151 in genuine news and 108 in misleading news. Figure 3.3 illustrates the distribution of emotional words in the two datasets. Considering the low figures obtained with the semantic attributes, the researcher concluded that emotion and informality are not relevant features for fake news detection in any of the datasets. Given that the slang and emotion lexicons were not extensive enough to cover the complete range of tokens in the two datasets, the arrays of the different measures had a comparatively significant proportion of zero values. Consequently, a more comprehensive list may lead to different results for the emotion and informality measures.

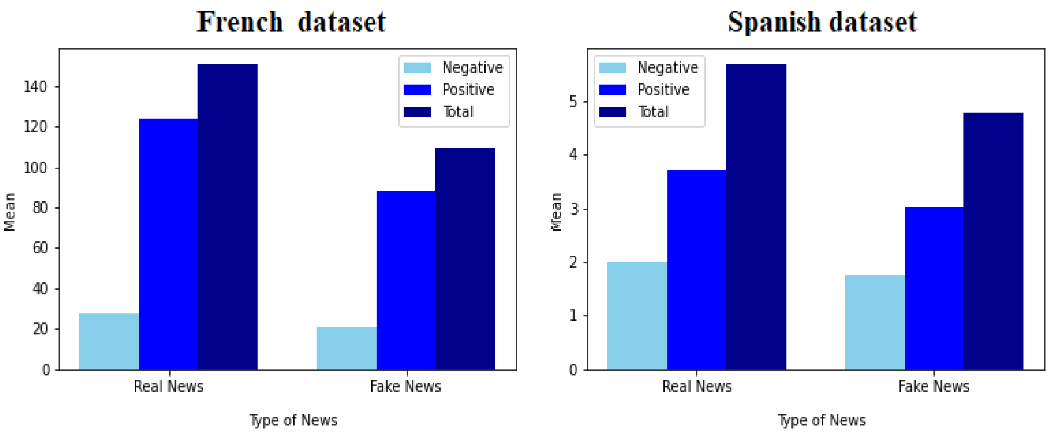


Figure 3.2: Mean of emotional words in the French and Spanish datasets

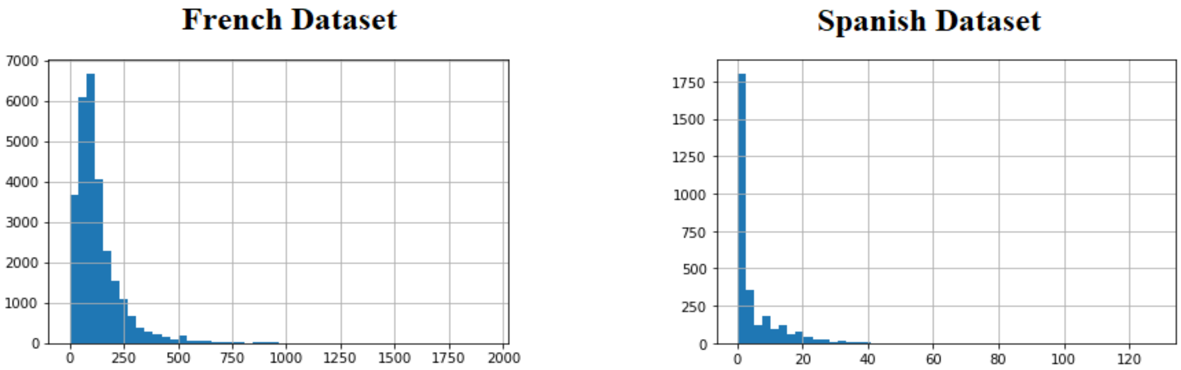


Figure 3.3: Illustration of the distribution of emotional content in both corpora

In relation to the lexical features considered, the model reported 77.7% accuracy with the character-n-gram level and word-n-gram-level TF-IDF models in the Spanish dataset, and 92% in the French corpus with the TF-IDF models. Figure 3.4 depicts the confusion matrices of the word-n-gram-level TF-IDF in both datasets after considering the lexical features. As it can be noticed, the number of FN errors has significantly decreased from 28 to 21 in the Spanish dataset, whereas the number of FN errors has increased from 27 to 32. Contrarily, the number of FP increased from 47 to 66 in the French dataset. The mean calculation of the different lexical features (see Table 3.2) in both datasets, demonstrates that every suggested lexical feature is to be considered distinguishing. The interpretation of these results may imply that while real news articles are frequently more extensive (larger number of words), they also tend to use punctuation more frequently to avoid long and vague sentences. Moreover, the frequent use of uppercase words may suggest that real news includes more Named Entities and factual data (such as organizations, proper names, places, dates).

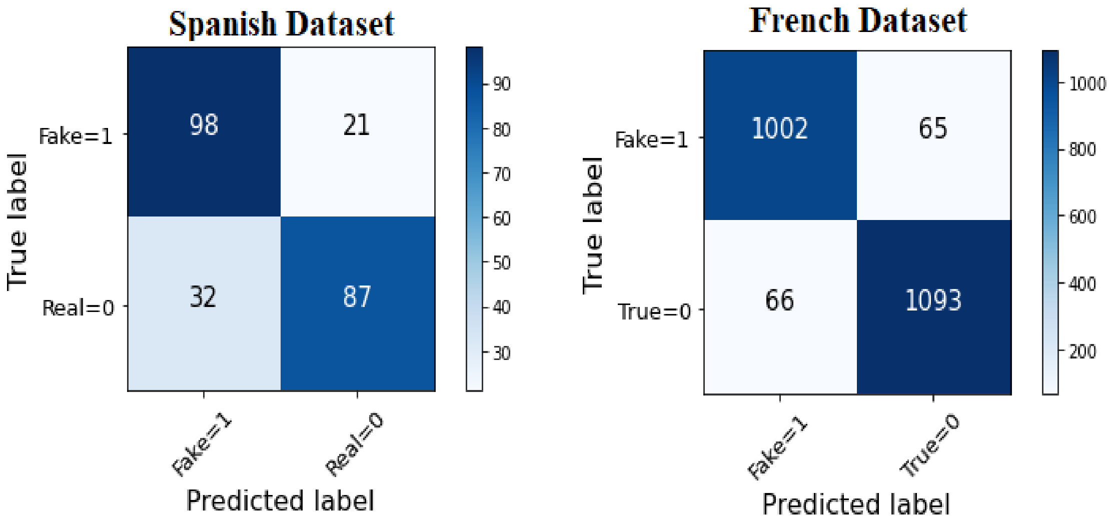


Figure 3.4: Confusion matrix of the word-n-gram-level TF-IDF models including lexical features

	French		Spanish	
	Fake	**Real**	**Fake**	**Real**
Character count	1340.607579	1942.946733	504.69527	758.64118
Word count	178.258784	253.364393	67.623649	98.608317
Title word count	33.213295	54.172181	14.356757	23.987928
Uppercase word count	3.215020	5.217346	3.643919	4.478203
Punctuation	46.133430	91.891815	17.895946	28.665996
Word density	7.492677	7.616330	7.166282	7.744124

Table 3.2: Comparison of the mean obtained with the selected lexical features

Despite the statistical significance of the hand-crafted lists of features, classification based on the aforesaid linguistic features was not successful. Overall, it is evident that TF-IDF models better results in the Spanish dataset. This outcome is related to how the TF-IDF model takes into account the occurrence of a word in the entire corpus and not in a single document. Accordingly, the same words in all documents are connected and affect the decision and rules of each classifier differently, while in Bag of words the words are treated individually in each document. On the contrary, the Bag of words models recorded higher accuracy with the French corpus.

2 Neural-network models

Classification results demonstrate the significant difference in the models' performance when trained in French or Spanish, which is mainly attributable to the number of samples available for each language. Regarding the French dataset, the testing accuracy for the first sequential LSTM model (with Count Vectorizer) was 94.4%, which did not outperform the accuracy of the baseline Count Vectorizer model (95%). Although the BLSTM model recorded 96% accuracy, the CNN+BLSTM outperformed all previous models, resulting in an accuracy of 98.6% on the test set. On the contrary, all the neural network-based models trained with the Spanish dataset resulted in overfitting, due to the limited vocabulary size, scarce number of samples available for training, and unbalanced nature of the corpus. When a model is overfitted, it implies that it is struggling to predict accurately on the unseen test set; the loss value does not steadily decrease, which implies that the model might not be learning at all. Misclassification is a concern that should be taken into consideration before making interpretations. Ignorance of the fact that the models are not perfectly accurate and can sometimes lead to misinterpretation of information, altering credibility. Some hypotheses can be made on why identical models yield significant results on the French dataset, while struggling to learn when dealing with the Spanish dataset. One aspect to be highlighted is the presence of different writing styles between reliable and misleading news articles. Another aspect to be considered is the unbalanced nature of the articles' length in the Spanish dataset. Although attempts are made to avoid overfitting, generalization can be challenging sometimes.

The researcher deployed different techniques to overcome overfitting and improve generalization on the Spanish dataset: early stopping during the training phase, dropout and Ridge Regularization (l2 = 0.01). Regularization is a technique to constrain the neural network from learning a model that is overfitted. In L2 regularization, a penalty term can be added to the cost function to push the estimated coefficients towards zero (but not zero), and not take more extreme values. L2 regularization was added to the dense layers. Dropout is another regularization technique in which randomly selected neurons are ignored during training with a set probability; by dropping some nodes the model tends to generalize better. Using dropout, the researcher can reduce interdependent learning among units, which may have led to overfitting. The researcher applied from 0.1 to 0.3 dropout to dense layers in the network. Additionally, early stopping, from the Keras library, is implemented in the models (tf.keras.callbacks.EarlyStopping). Finally, the learning rate was reduced to 0.0001. After applying these techniques, the researcher observed that the regularization strategies employed did not significantly improve the performance of any of the models, with the validation loss as the quantity to be monitored and patience of 5. After the implementation of these techniques, the LSTM model (with Count Vectorizer) recorded the highest accuracy of 76.6%, followed by the CNN+BLSTM (76%) and BLSTM (75.8%). The evaluation metrics employed in this study are: precision, recall, f1-score and accuracy. The results were analysed according to the category 'Fake' (1). Tables 3.3 and 3.4 illustrate the classification report of each dataset to be compared.

French Dataset		**Test Set**			
Baseline Models	Features	**A**	**P**	**R**	**F1**
CV	No applicable	95.01	94	95	94.7
CV	Slang + Emotion	95.06	94	95	94.8
CV	Lexical Features	92	89	94	91.7
Word-level TF-IDF	No applicable	94.8	94.8	94	94
Word-level TF-IDF	Slang + Emotion	94.8	94.8	94	94
Word-level TF-IDF	Lexical Features	91	98	92	91
Word-n-gram level TF-IDF	No applicable	90	90	90	90
Word-n-gram level TF-IDF	Slang + Emotion	90	90	90	90
Word-n-gram level TF-IDF	Lexical Features	91	89	92	91
Char.-n-gram level TF-IDF	No applicable	92	91	92	92
Char.-n-gram level TF-IDF	Slang + Emotion	92	91.9	92	92
Char.-n-gram level TF-IDF	Lexical Features	91	89	92	91
DL Models	Features				
LSTM	Word Vectors	94.4	92	97	95
BLSTM	Word embeddings	96	97	95	96
CNN+BLSTM	Pre-trained word embeddings	98.6	99	99	99

Table 3.3: Classification results using ML and DL-based models in the French dataset

Spanish Dataset		**Test Set**			
Baseline Models	Features	**A**	**P**	**R**	**F1**
CV	No applicable	69.7	67	70	68.9
CV	Slang + Emotion	69.7	67	70.8	68.9
CV	Lexical Features	70.5	77.3	68	72
Word-level TF-IDF	No applicable	76.8	76.4	77	76.7
Word-level TF-IDF	Slang + Emotion	76.8	76	77	76.7
Word-level TF-IDF	Lexical Features	77.7	82	75	78.7
Word-n-gram level TF-IDF	No applicable	62.6	57.9	63	60.5
Word-n-gram level TF-IDF	Slang + Emotion	62.6	57.9	63.8	60.7
Word-n-gram level TF-IDF	Lexical Features	77.7	82	75	78.7
Char.-n-gram level TF-IDF	No applicable	74	72	75	73.8
Char.-n-gram level TF-IDF	Slang + Emotion	74.3	72	75	73.8
Char.-n-gram level TF-IDF	Lexical Features	77.7	82	75	78.7
DL Models	Features				
LSTM	Word Vectors	76.6	76	78	77
BLSTM	Word embeddings	75.8	75	76	75
CNN+BLSTM	Pre-trained word embeddings	76	77	77	77

Table 3.4: Classification results using ML and DL-based models in the Spanish dataset

Given the unsatisfying results obtained with the Spanish dataset, one of the strategies followed during experiments, consisted of evaluating the capacity of the already described DL models trained with an augmented Spanish corpus. This new corpus contained 6,000 articles (3,000 True, 3,000 Fake), which comprised the original 2,971 samples of the Spanish dataset together with 3,029 news articles translated from the French dataset into Spanish using the Google translation API. The articles of the new dataset were cleaned and pre-processed according to the steps described in Section 2 (Chapter 2). The maximum length was set to 2,806. The results demonstrated superior performance in the models when trained with the augmented dataset. The highest accuracy was obtained with the CNN+BLSTM model (accounting for 98.7% accuracy), followed by the LSTM (85.8%), and BLSTM (84.8%) models. This experiment proves the models' high performance to automatically detect fake news, justifying the previous models' poor performance, in the Spanish dataset, due to the limited number of samples and the unbalanced nature of the articles.

Augmented Spanish Dataset		**Test Set**			
DL Models	Features	A	P	R	F1
LSTM	Word vectors	85.8	85	87	86
BLSTM	Word embeddings	84.8	88	81	84
CNN+BLSTM	Pre-trained word embeddings	98.7	99	96	97

Table 3.5: Classification results of Deep Learning-based models in the Augmented Spanish dataset

3 Conclusion

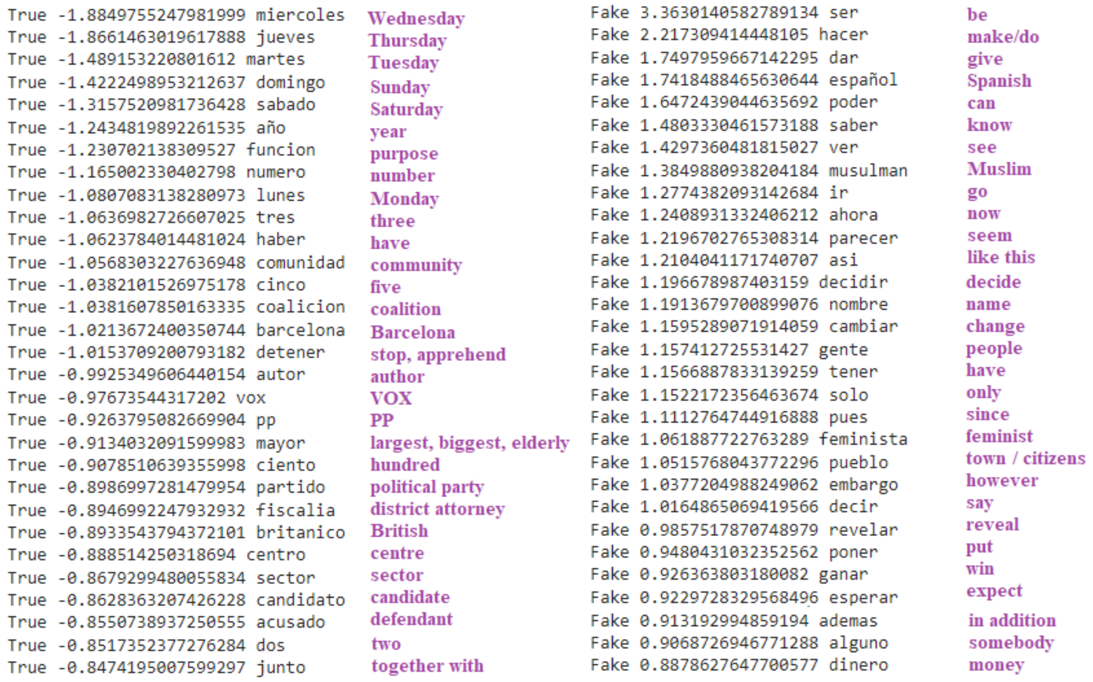
Despite their benefits, deep learning models have several practical limitations, such as the difficulty in determining the optimal hyper-parameters for each problem and dataset, the need for large training datasets, the lack of interpretability, and the computational complexity and time required, all of which have a direct impact on the DL models' performance. After discovering the optimal values for each parameter on all classifiers, the CNN+BLSTM model recorded the highest accuracy of over 98% on both datasets. The results also demonstrated that the LSTM model did not outperform the accuracy of the Logistic Regression baseline model. The experiments conducted with the augmented data demonstrated that more training data leads to better performance by the deep learning models. In conclusion, the capability of automatic feature extraction with deep learning models has proved to be paramount in the accurate detection of fake news.

Chapter 4

Findings and Discussion

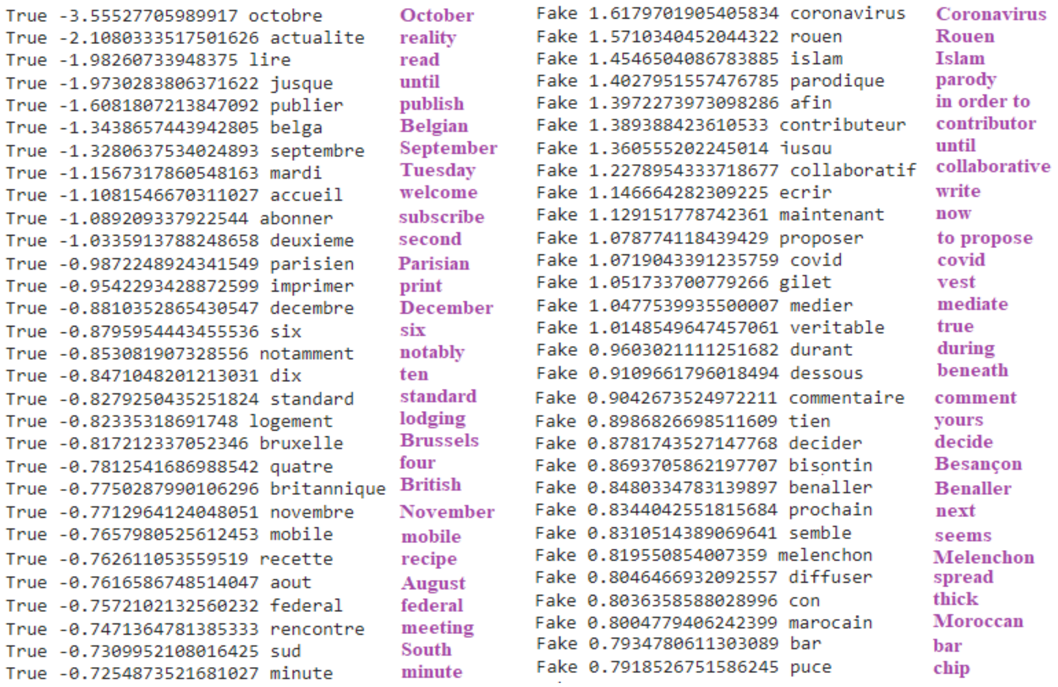
1 Error Analysis at the word level

The researcher attempted to analyse the most significant vectors for the baseline classification models which recorded the highest accuracy: the CountVectorizer (French dataset) and word-level TF-IDF (Spanish dataset). This was performed by compressing the classifiers' coefficients with the features names, and then assessing the top forty ranking index‘s values for real and fake classes. Figure 4.1 illustrates a sampling of the "most fake" and "most real" features produced by the word-level TF-IDF in the Spanish dataset. Similarly, Figure 4.2 depicts the most important words in fake and real news articles provided by Count Vectorizer in the French dataset.



```
True -1.8849755247981999 miercoles  Wednesday                 Fake 3.3630140582789134 ser           be
True -1.8661463019617888 jueves     Thursday                  Fake 2.217309414448105 hacer          make/do
True -1.489153220801612 martes      Tuesday                   Fake 1.7497959667142295 dar           give
True -1.4222498953212637 domingo    Sunday                    Fake 1.7418488465630644 español       Spanish
True -1.3157520981736428 sabado     Saturday                  Fake 1.6472439044635692 poder         can
True -1.2434819892261535 año        year                      Fake 1.4803330461573188 saber         know
True -1.230702138309527 funcion     purpose                   Fake 1.4297360481815027 ver           see
True -1.165002330402798 numero      number                    Fake 1.3849880938204184 musulman      Muslim
True -1.0807083138280973 lunes      Monday                    Fake 1.2774382093142684 ir            go
True -1.0636982726607025 tres       three                     Fake 1.2408931332406212 ahora         now
True -1.0623784014481024 haber      have                      Fake 1.2196702765308314 parecer       seem
True -1.0568303227636948 comunidad  community                 Fake 1.2104041171740707 asi           like this
True -1.0382101526975178 cinco      five                      Fake 1.196678987403159 decidir        decide
True -1.0381607850163335 coalicion  coalition                 Fake 1.1913679700899076 nombre        name
True -1.0213672400350744 barcelona  Barcelona                 Fake 1.1595289071914059 cambiar       change
True -1.0153709200793182 detener    stop, apprehend           Fake 1.157412725531427 gente          people
True -0.9925349606440154 autor      author                    Fake 1.1566887833139259 tener         have
True -0.97673544317202 vox          VOX                       Fake 1.1522172356463674 solo          only
True -0.9263795082669904 pp         PP                        Fake 1.1112764744916888 pues          since
True -0.9134032091599983 mayor      largest, biggest, elderly Fake 1.061887722763289 feminista      feminist
True -0.9078510639355998 ciento     hundred                   Fake 1.0515768043772296 pueblo        town / citizens
True -0.8986997281479954 partido    political party           Fake 1.0377204988249062 embargo       however
True -0.8946992247932932 fiscalia   district attorney         Fake 1.0164865069419566 decir         say
True -0.8933543794372101 britanico  British                   Fake 0.9857517870748979 revelar       reveal
True -0.888514250318694 centro      centre                    Fake 0.9480431032352562 poner         put
True -0.8679299480055834 sector     sector                    Fake 0.926363803180082 ganar          win
True -0.8628363207426228 candidato  candidate                 Fake 0.9229728329568496 esperar       expect
True -0.8550738937250555 acusado    defendant                 Fake 0.913192994859194 ademas         in addition
True -0.8517352377276284 dos        two                       Fake 0.9068726946771288 alguno        somebody
True -0.8474195007599297 junto      together with             Fake 0.8878627647700577 dinero        money
```

Figure 4.1: Most real and fake features produced by word n-gram-level TF-IDF: Spanish dataset



```
True -3.55527705989917 octobre        October       Fake 1.6179701905405834 coronavirus   Coronavirus
True -2.1080333517501626 actualite    reality       Fake 1.5710340452044322 rouen         Rouen
True -1.98260733948375 lire           read          Fake 1.4546504086783885 islam         Islam
True -1.9730283806371622 jusque       until         Fake 1.4027951557476785 parodique     parody
True -1.6081807213847092 publier      publish       Fake 1.3972273973098286 afin          in order to
True -1.3438657443942805 belga        Belgian       Fake 1.389388423610533 contributeur   contributor
True -1.3280637534024893 septembre    September     Fake 1.360555202245014 iusau          until
True -1.1567317860548163 mardi        Tuesday       Fake 1.2278954333718677 collaboratif  collaborative
True -1.1081546670311027 accueil      welcome       Fake 1.146664282309225 ecrir          write
True -1.089209337922544 abonner       subscribe     Fake 1.129151778742361 maintenant     now
True -1.0335913788248658 deuxieme     second        Fake 1.078774118439429 proposer       to propose
True -0.9872248924341549 parisien     Parisian      Fake 1.0719043391235759 covid         covid
True -0.9542293428872599 imprimer     print         Fake 1.051733700779266 gilet          vest
True -0.8810352865430547 decembre     December      Fake 1.0477539935500007 medier        mediate
True -0.8795954443455536 six          six           Fake 1.0148549647457061 veritable     true
True -0.853081907328556 notamment     notably       Fake 0.9603021111251682 durant        during
True -0.8471048201213031 dix          ten           Fake 0.9109661796018494 dessous       beneath
True -0.8279250435251824 standard     standard      Fake 0.9042673524972211 commentaire   comment
True -0.82335318691748 logement       lodging       Fake 0.8986826698511609 tien          yours
True -0.817212337052346 bruxelle      Brussels      Fake 0.8781743527147768 decider       decide
True -0.7812541686988542 quatre       four          Fake 0.8693705862197707 bisontin      Besançon
True -0.7750287990106296 britannique  British       Fake 0.8480334783139897 benaller      Benaller
True -0.7712964124048051 novembre     November      Fake 0.8344042551815684 prochain      next
True -0.7657980525612453 mobile       mobile        Fake 0.8310514389069641 semble        seems
True -0.762611053559519 recette       recipe        Fake 0.819550854007359 melenchon      Melenchon
True -0.7616586748514047 aout         August        Fake 0.8046466932092557 diffuser      spread
True -0.7572102132560232 federal      federal       Fake 0.8036358588028996 con           thick
True -0.7471364781385333 rencontre    meeting       Fake 0.8004779406242399 marocain      Moroccan
True -0.7309952108016425 sud          South         Fake 0.7934780611303089 bar           bar
True -0.7254873521681027 minute       minute        Fake 0.7918526751586245 puce          chip
```

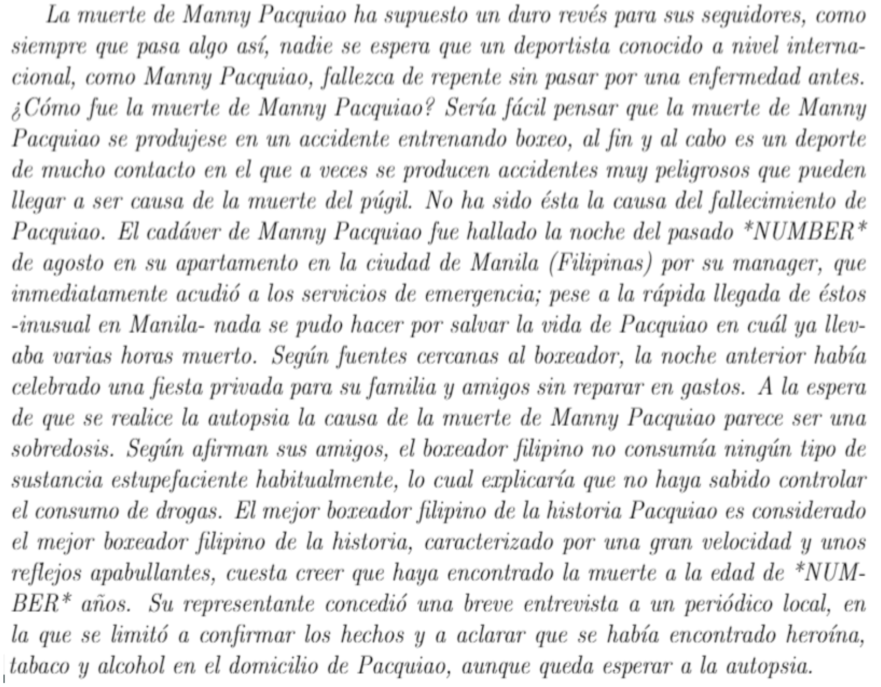
Figure 4.2: Most real and fake features produced by word CV: French dataset

A number of observations were made on the most prominent feature names. According to the above picture, time-related words have been reported to contribute more with respect to the real class in both datasets. Examples of words that provide temporal content are: "miércoles" (Wednesday) and "año" (year) in Spanish; "October" (October) and "Mardi" (Tuesday) in French. The use of numerical and quantity figures is also a common pattern to identify real news in both datasets (e.g., "cinco" (five) or "ciento" (hundred) in the Spanish dataset; as well as "deuxième" (second) or "quatre" (four) in the French dataset. Named Entity words are also recurrent in genuine articles. Instances of named entities are: proper names ('Melenchon'), city names ("Barcelona","Bruxelle" (Brussels) nationalities (e.g., French word "Parisien" (Parisian)) as well as organizational entities (e.g. "VOX" or "PP", which are Spanish political parties).

Regarding the most prominent featuring vectors in fake articles, it could be highlighted the presence of subjective words, such as "parecer" in Spanish, and "semble" in French, both meaning "seem". Another observation to be made in 'highly fake' words, is related to the paradoxical use of the word 'truth' ("veritable" in French), as an attempt to convince the reader about the veracity of the pieces of news. The researcher also observed that trending topics are more frequent in fake news articles, examples of these topics are: coronavirus, feminism, and Islamism. Another paramount aspect in the identification of misleading articles is the presence of countless nouns, such as "gente" (people), "alguno" (somebody) and "pueblo" (citizens) in the Spanish dataset. Generally, words that are characterized by a formally formatted style and objectiveness are mostly observed to be in real data records; whereas subjective and abstract words appear more occasionally on fake news articles.

2 Error Analysis at the document level

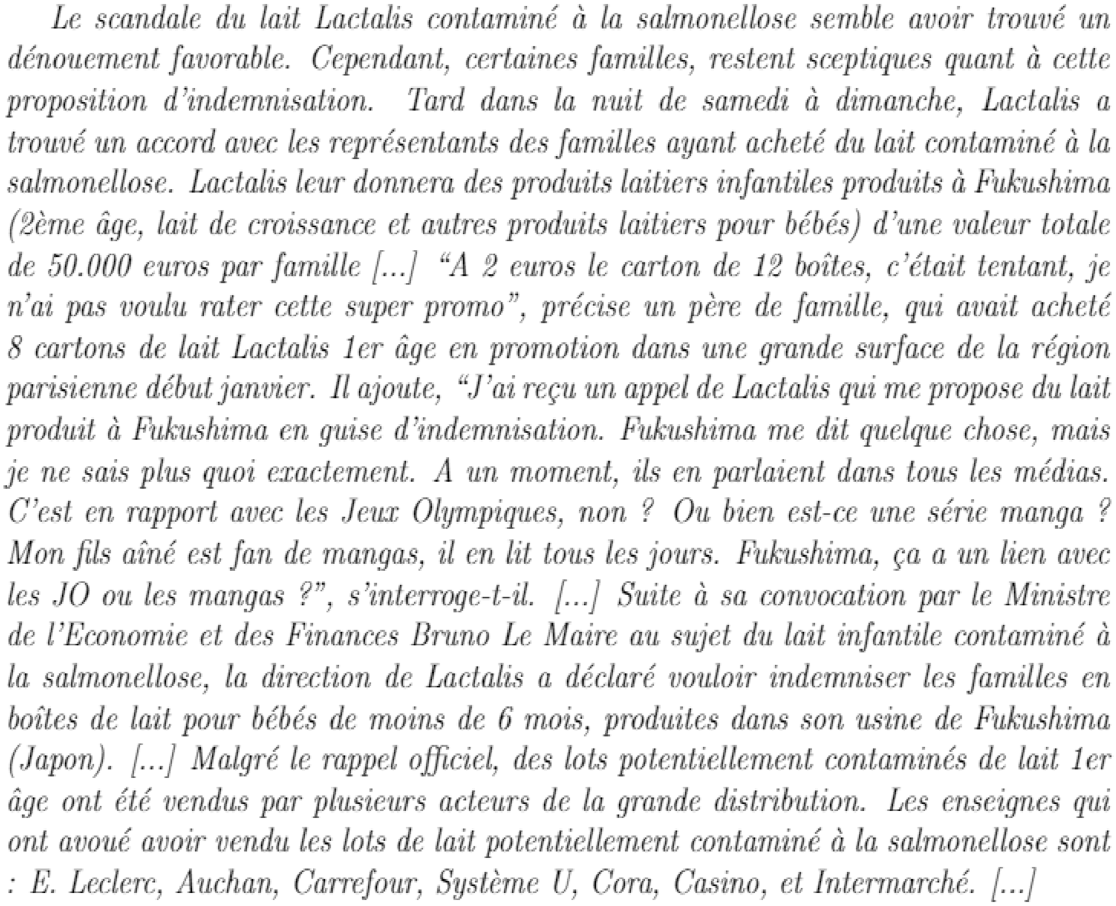
As illustrated by the confusion matrices in Figure 3.1 and, the best performing model in both datasets mainly fail with the False Negatives, as it mistakenly assigns 'real' labels to fake articles. In the Spanish dataset, the word-n-gram-level TF-IDF model misclassified 28 FN; whereas the CV implemented in the French dataset misclassified 64 FN. Aiming to obtain a deeper understanding of the baseline classification results, some FP and FN articles were extracted and analysed from each classifier. A rigorous analysis was conducted to discover the possible linguistic patterns in the misclassified False Negative articles produced by the best-performing systems. The instances analysed in the Spanish dataset revealed that the model mainly failed at classifying articles belonging to the dataset downloaded from Kaggle, as it consists of very short sentences and there is not sufficient information to correctly label the articles. The most common lexical features of the FN instances are the presence of quotations, rhetorical questions as well as subjective and informal words. Figure 4.3 shows an example of a Spanish fake article misclassified for 'real'. This article relates the death of a world-renowned Filipino boxer, Manny Pacquiao, who presumably had died of an overdose. As it can be observed, the article contains indefinite words (e.g. "nadie" (nobody), "siempre" (always)), as well as emotional adjectives ("duro" (hard), "peligroso" (dangerous)). Another linguistic characteristic to be highlighted is the use of rhetorical questions ("¿Cómo fue la muerte de Manny Pacquiao?"('How was the death of Manny Pacquiao?)). Given the fact that most of the misclassified fake news articles imitate real pieces of news, there is not any significant linguistic feature that aids the model or the researcher to classify these articles with complete certainty.



La muerte de Manny Pacquiao ha supuesto un duro revés para sus seguidores, como siempre que pasa algo así, nadie se espera que un deportista conocido a nivel internacional, como Manny Pacquiao, fallezca de repente sin pasar por una enfermedad antes. ¿Cómo fue la muerte de Manny Pacquiao? Sería fácil pensar que la muerte de Manny Pacquiao se produjese en un accidente entrenando boxeo, al fin y al cabo es un deporte de mucho contacto en el que a veces se producen accidentes muy peligrosos que pueden llegar a ser causa de la muerte del púgil. No ha sido ésta la causa del fallecimiento de Pacquiao. El cadáver de Manny Pacquiao fue hallado la noche del pasado *NUMBER* de agosto en su apartamento en la ciudad de Manila (Filipinas) por su manager, que inmediatamente acudió a los servicios de emergencia; pese a la rápida llegada de éstos -inusual en Manila- nada se pudo hacer por salvar la vida de Pacquiao en cuál ya llevaba varias horas muerto. Según fuentes cercanas al boxeador, la noche anterior había celebrado una fiesta privada para su familia y amigos sin reparar en gastos. A la espera de que se realice la autopsia la causa de la muerte de Manny Pacquiao parece ser una sobredosis. Según afirman sus amigos, el boxeador filipino no consumía ningún tipo de sustancia estupefaciente habitualmente, lo cual explicaría que no haya sabido controlar el consumo de drogas. El mejor boxeador filipino de la historia Pacquiao es considerado el mejor boxeador filipino de la historia, caracterizado por una gran velocidad y unos reflejos apabullantes, cuesta creer que haya encontrado la muerte a la edad de *NUMBER* años. Su representante concedió una breve entrevista a un periódico local, en la que se limitó a confirmar los hechos y a aclarar que se había encontrado heroína, tabaco y alcohol en el domicilio de Pacquiao, aunque queda esperar a la autopsia.

Figure 4.3: Spanish fake article classified as real

By observing the French articles, the researcher discovered some linguistic aspects which are present in the misclassified documents. Figure 4.4 shows an instance of a French fake article misclassified for ‘real’. This piece of news provides fake information about alleged milk contaminated with salmonellosis. Similar to the Spanish misclassified articles, the researcher observed a tendency for quotation marks and rhetorical questions (“C’est en rapport avec les Jeux Olympiques, non?” (It’s related to the Olympics, right?)). Another feature which is frequent in these misclassified fake articles is the use of a limited vocabulary with several repetitions: “familles” (families), “indemniser” (indemnify). The scarce use of synonyms highlights the unreliability of the news story. In addition, the researcher recognized the use of phrases carrying high sentiment adjectives and intensifying adverbs such as “favorable” (favourable), “potentiellement” (potentially). On the contrary, the inclusion of figures related to price (50,000 euros), time (6 “mois” (months)), and quantity (12 “boîtes” (bottles)), as well as the reference of named entities (“ministre de l’economie et des finances” (Minister of Economy and Finance); Bruno Le Maire (proper name)) confuses the reader and even the model to correctly classify similar articles as “fake”.



Le scandale du lait Lactalis contaminé à la salmonellose semble avoir trouvé un dénouement favorable. Cependant, certaines familles, restent sceptiques quant à cette proposition d’indemnisation. Tard dans la nuit de samedi à dimanche, Lactalis a trouvé un accord avec les représentants des familles ayant acheté du lait contaminé à la salmonellose. Lactalis leur donnera des produits laitiers infantiles produits à Fukushima (2ème âge, lait de croissance et autres produits laitiers pour bébés) d’une valeur totale de 50.000 euros par famille [...] “A 2 euros le carton de 12 boîtes, c’était tentant, je n’ai pas voulu rater cette super promo”, précise un père de famille, qui avait acheté 8 cartons de lait Lactalis 1er âge en promotion dans une grande surface de la région parisienne début janvier. Il ajoute, “J’ai reçu un appel de Lactalis qui me propose du lait produit à Fukushima en guise d’indemnisation. Fukushima me dit quelque chose, mais je ne sais plus quoi exactement. A un moment, ils en parlaient dans tous les médias. C’est en rapport avec les Jeux Olympiques, non ? Ou bien est-ce une série manga ? Mon fils aîné est fan de mangas, il en lit tous les jours. Fukushima, ça a un lien avec les JO ou les mangas ?”, s’interroge-t-il. [...] Suite à sa convocation par le Ministre de l’Economie et des Finances Bruno Le Maire au sujet du lait infantile contaminé à la salmonellose, la direction de Lactalis a déclaré vouloir indemniser les familles en boîtes de lait pour bébés de moins de 6 mois, produites dans son usine de Fukushima (Japon). [...] Malgré le rappel officiel, des lots potentiellement contaminés de lait 1er âge ont été vendus par plusieurs acteurs de la grande distribution. Les enseignes qui ont avoué avoir vendu les lots de lait potentiellement contaminé à la salmonellose sont : E. Leclerc, Auchan, Carrefour, Système U, Cora, Casino, et Intermarché. [...]

Figure 4.4: French fake article classified as real

3 Learning curve results for CNN+BLSTM with trainable embeddings

Aiming to investigate whether large amounts of training data are required for the identification of fake content, the researcher analyzed the learning trend of the CNN+BLSTM models in the French and augmented Spanish datasets. Figures 4.5 and 4.6 show the plotting of the learning curve for CNN+BLSTM with trainable embedding using incremental amounts of data. Six experiments were run in the agumented Spanish dataset, adding 1000, 2000, 3000, 4000, 5000 and 6000 samples to the training set. Similarly, 5,000, 10,000, 15,000, 20,00 and 27,816 samples were added to the training set in the French dataset. Overall, the learning trend depicts a relatively high accuracy with the first number of articles. The learning curve quickly reaches a plateau of around 96.7% and 86% accuracy in the French and Augmented Spanish datasets, respectively. The incremental classification suggests that larger quantities of training data would improve the classification performance in the Spanish dataset.

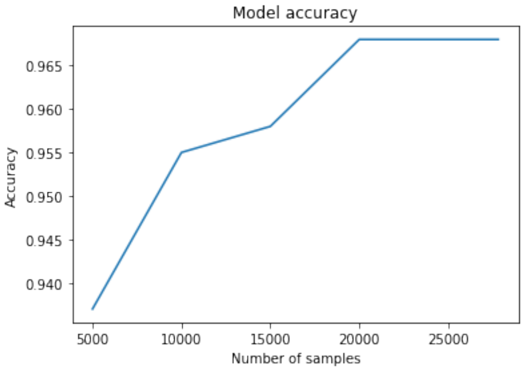


Figure 4.5: Learning curve results for CNN+BLSTM with trainable embedding: French dataset

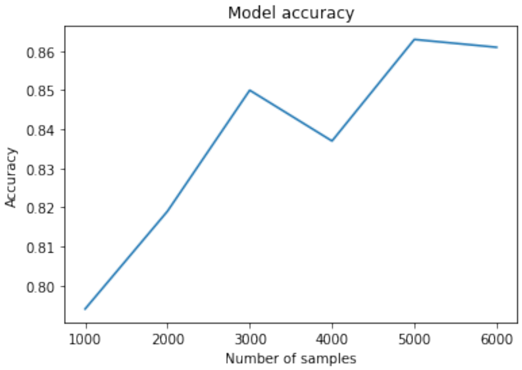


Figure 4.6: Learning curve results for CNN+BLSTM with trainable embedding: Spanish dataset

4 Comparison to previous research of the results obtained with the original Spanish corpus

Martínez-Gallego et al. (2021) implemented a similar study by combining the same two Spanish Datasets: The Spanish Fake News Corpus (Posadas-Durán et al., 2019), and 'Fake News in Spanish'[1] downloaded from Kaggle. The authors also removed stop words, but applied stemming instead of lemmatization. The ML algorithms chosen by the researchers were: SVM, RF, Gradient Boosting Tree (GBT), and Multi-Layer Perceptron (MLP). For the case of DL classifiers, two types of layers were used by the researchers: LSTM-RNN, and CNN. The Deep Learning architectures were built on top of different pre-trained Word Embedding representations, including Global Vectors for Word Representation (GloVe), Embeddings from a Language Model (ELMo), BERT, and BETO (a BERT version trained on a large corpus in Spanish). According to the results, the highest accuracy of 80% was obtained with RF estimator using TF-IDF text representation: it outperformed by 3% the results obtained in this thesis with the TF-IDF Logistic Regression model. Despite the fact that the combination of pre-trained BETO and LSTM also yield 80% accuracy, the authors concluded that the models exhibited a trend of overfitting due to the limited number of samples available for training.

Similar to the results achieved in this thesis, the authors observed that the regularization strategies employed in their project did not significantly improve the performance of their models (dropout, lasso regularization). Another strategy followed by Martínez-Gallego et al. (2021) during their experiments consisted of evaluating the capacity of a DL model trained with an English corpus to predict fake Spanish news translated into English using the Google translation API. Therefore, the researchers trained the models with the dataset in English and validated with the translated dataset. When validating the models using the translated dataset, the BERT embedding with CNN recorded 53% accuracy, whereas the ElMo Embedding with CNN achieved 52.5%. Different from the authors' approach, this research attempted to use Google translation API to expand the Spanish corpus by adding 2,000 additional samples extracted from the French dataset and translated into Spanish. Our Neural-network CNN+BLSTM model outperformed their results with up to 98% in the augmented Spanish dataset.

[1] https://www.kaggle.com/arseniitretiakov/noticias-falsas-en-espaol

Chapter 5

Conclusion

1 Problem Evaluation

This thesis proposed a novel approach for the classification of deliberate fake news in Spanish and French, implementing various state-of-the-art deep learning and baseline models. Experiments revealed that a proposed architecture combining CNN and BLSTM with pre-trained word embeddings outperformed other proposed models. The CNN+LSTM architecture successfully identified Spanish and French fake news with 98% accuracy. This study demonstrates that machine learning has the capability to learn subtle linguistic patterns that are difficult for humans to identify.

A number of observations should be considered. Firstly, the obtained results should be analysed according to the consequences addressed in this task. Since misclassified fake news articles may have devastating implications, it is not reasonable to assign an equal cost to the false negatives and false positives. Type II error is more costly and consequential than a Type I error, that is to say, mistakenly taking deceptive news for real far exceeds the cost of dismissing a piece of real news as fake. Accordingly, an automatic detection with a high recall is fundamental in preventing the spread of the type of misinformation included in the positive dataset of the current research. Secondly, the experiments demonstrated that the CNN+BLSTM neural architecture with pre-trained word embeddings as input is sufficient to capture fake news without the aid of any hand-crafted linguistic features. Overall, the use of artificial neural networks shows potential in fake news detection.

2 Limitations

The limited research and publicly available datasets for deliberate fake news in French and Spanish constituted a challenge. Although all existing datasets were merged in an attempt to have a viable volume of data, the Spanish datasets employed were yet not sufficient to train neural networks. Despite all the techniques applied to solve overfitting (regularization techniques, artificial extension of the data) in the Spanish corpus, no viable solutions were achieved with the original corpus. Nonetheless, the augmentation of the Spanish dataset with the translated French articles yields significant results. The researcher attempted to apply some of the procedures used in the literature of English fake news detection, however many of them were not feasible with the Spanish and French datasets. For example, a considerable number of studies showed that fake and real news articles are notably distinguishable, specifically in the title of the articles. While fake news titles use significantly fewer stop-words and nouns, they include more proper nouns and verb phrases. Because most articles did not contain their respective titles, this approach was not possible. Another paramount challenge was related to the emotion and informality measurement; the vast majority of the articles rendered a zero value. Hence, a more extensive and elaborated dictionary should be compiled. The researcher was unable to provide an adequate statistical measure of the informality typical of fake news due to a large number of irrelevant results.

3 Potential Further Research

Regardless of the abundant academic research addressing fake news detection, the area has still the potential for improvement, innovation and experimentation; seeking insights on the nature of deliberate fake news may lead to more efficient and accurate models. This study provides evidence that AI technologies can indeed automatically prevent the early spread of misleading information. A further step would be to build a software application to detect fake news in French and Spanish. Another aspect in which this project could be expanded is by building a viable Spanish dataset that could be used to develop more robust systems from the best strategy the researcher found (CNN+BLSTM). More experiments combining hyperparameter values and network architectures could also be implemented. Another aspect in which this project could be expanded is by comparing the models' performance to humans in labelling articles as fake or real. Comparing the accuracies would be beneficial in deciding whether or not the model outperforms the human abilities to separate fake from real news articles. If humans are more accurate than the model, it may mean that the researcher needs to choose more deceptive fake news examples.

Appendix A

Declaration

UNIVERSITY OF WOLVERHAMPTON

SCHOOL OF HUMANITIES : MA in Computational Linguistics

Name: Isabel Olmos Canovas

Date: 31-December-2021

Title: Is this really fake?: A study of Spanish and French Automatic Fake News Detection

Module Code: Dissertation 7LN007/UM1

Presented in partial fulfilment of the assessment requirements for the above award.

Supervisors:

Dr. Hadeel Saadany & Dr. Rocio de Caro

Declaration:

(i) This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

(ii) It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the University a perpetual royalty-free licence to do all or any of those things referred to in section 16(i) of the Copyright Designs and Patents Act 1988 (viz: to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make adaptation of the work). (iii) This project did not involve direct contact with human subjects, and hence did not require approval from the LSSC Ethics Committee.

Date: 29/12/2021

Student's signature: Isabel Olmos Cánovas

Bibliography

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.

Costel-Sergiu Atodiresei, Alexandru Tănăselea, and Adrian Iftene. Identifying fake news and fake users on twitter. *Procedia Computer Science*, 126:451–461, 2018.

Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, 41(3):430–454, 2014.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE, 2019.

Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164, 2009.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

Anshika Choudhary and Anuja Arora. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171, 2021.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.

Tim Dwyer. Media manipulation, fake news, and misinformation in the asia-pacific region. *Journal of Contemporary Eastern Asia*, 18(2):9–15, 2019.

Agata Giełczyk, Rafał Wawrzyniak, and Michał Choraś. Evaluation of the existing tools for fake news detection. In *IFIP International Conference on Computer Information Systems and Industrial Management*, pages 144–151. Springer, 2019.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213, 2019.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer, 2005.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

Mauricio Gruppi, Benjamin D Horne, and Sibel Adali. An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*, 2018.

Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. Multilingual fake news detection with satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*, 2019.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789*, 2020.

Xixuan Huang, Jieying Xiong, and Shengyi Jiang. Gduf_dm at fakedes 2021: Spanish fake news detection with bert and sample memory. *Education*, 6(9):4, 2021.

Marcia K Johnson and Carol L Raye. Reality monitoring. *Psychological review*, 88(1): 67, 1981.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44, 2020.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788, 2021.

Junaed Younus Khan, Md Khondaker, Tawkat Islam, Anindya Iqbal, and Sadia Afroz. A benchmark study on machine learning methods for fake news detection. *arXiv preprint arXiv:1905.04749*, 2019.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*, 2018.

Zhan Liu, Shaban Shabani, Nicole Glassey Balet, and Maria Sokhn. Detection of satiric news on social media: Analysis of the phenomenon with a french dataset. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6. IEEE, 2019.

Kevin Martínez-Gallego, Andrés M Álvarez-Ortiz, and Julián D Arias-Londoño. Fake news detection in spanish using deep learning techniques. *arXiv preprint arXiv:2110.06461*, 2021.

Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth international AAAI conference on web and social media*, 2015.

Sina Mohseni, Eric Ragan, and Xia Hu. Open issues in combating fake news: Interpretability as an opportunity. *arXiv preprint arXiv:1904.03016*, 2019.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*, 2018.

Özlem Özgöbek and Jon Atle Gulla. Towards an understanding of fake news. In *CEUR workshop proceedings*, volume 2041, pages 35–42, 2017.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876, 2019.

Kenneth Rapoza. Can 'fake news' impact the stock market? *Forbes News*, 2017.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. " everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. *arXiv preprint arXiv:1706.05924*, 2017.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.

Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 573–582, 2019.

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.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Trevino, and Juan Gordon. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence*, pages 1–14. Springer, 2012.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

Tina Esther Trueman, Ashok Kumar, P Narayanasamy, and J Vidya. Attention-based c-bilstm for fake news detection. *Applied Soft Computing*, page 107600, 2021.

Udo Undeutsch. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der psychologie*, 11:26–181, 1967.

Dinesh Kumar Vishwakarma and Chhavi Jain. Recent state-of-the-art of fake news detection: A review. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2020.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22, 2014.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, 2017.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*, 2018.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837, 2019.

Melissa Zimdars and Kembrew McLeod. *Fake news: understanding media and misinformation in the digital age.* MIT Press, 2020.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier, 1981.