FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages

Lucas Cabral[1], José Maria Monteiro[1], José Wellington Franco da Silva[1], César Lincoln Mattos[1]
and Pedro Jorge Chaves Mourão[2]

[1]*Computer Science Department, Federal University of Ceará, Fortaleza, Ceará, Brazil*

[2]*State University of Ceará, Fortaleza, Brazil*

Keywords: Misinformation Detection, Fake News Detection, Natural Language Processing, WhatsApp, Social Media.

Abstract: In the past few years, the large-scale dissemination of misinformation through social media has become a critical issue, harming the trustworthiness of legit information, social stability, democracy and public health. Thus, developing automated misinformation detection methods has become a field of high interests both in academia and in industry. In many developing countries such as Brazil, India, and Mexico, one of the primary sources of misinformation is the messaging application WhatsApp. Despite this scenario, due to the private messaging nature of WhatsApp, there still few methods of misinformation detection developed specifically for this platform. In this work we present the FakeWhatsApp.BR, a dataset of WhatsApp messages in Brazilian Portuguese, collected from Brazilian public groups and manually labeled. Besides, we evaluated a series of misinformation classifiers combining Natural Language Processing-based techniques of feature extraction and a set of well-know machine learning algorithms, totaling 108 different scenarios. Our best result achieved a F1 score of 0.73, and the analysis of errors indicates that they occur mainly due to the predominance of short texts that accompany media files. When texts with less than 50 words are filtered, the F1 score rises to 0.87.

1 INTRODUCTION

The rise of social media platforms revolutionized how we produce, share, and consume information, greatly improving its transmission velocity and available volume. The boundaries between information production and sharing are blurring fastly. However, while social networks made wider access to good information, its highly decentralized and unregulated environment allows the mass proliferation of misinformation (Vosoughi et al., 2018; Guo et al., 2019; Su et al., 2020). Through these platforms, misinformation can deceive thousands of people in a short time, bringing great harm to individuals, companies, or even society. Misinformation is a broad concept that can be defined as misrepresented information, including fabricated, misleading, false, fake, deceptive, or distorted information (Su et al., 2020). This comprehensive definition covers a variety of specific, and sometimes overlapping, types of such as fake news (Lazer et al., 2018), rumor(Shu et al., 2017), deception (Maalej, 2001) and hoaxes. In particular, the term fake news, despite specifically describe intentionally misleading information written as journalistic news, has become very present in popular culture and sometimes is informally used as a misinformation synonym.

Misinformation is usually created with malicious intentions to manipulate public opinion, harm individuals, organizations, or social groups, and obtain economic or political gains. Moreover, misinformation spreads faster, deeper, and broader in social media than legit information. Further, due to the high volume of information that we are exposed to when using social media, humans have a limited ability to distinguish true information from misinformation (Vosoughi et al., 2018; Qiu et al., 2017). The widespread of misinformation causes a major social problem, as breaks the trustworthiness of legit information, harming the democracy, justice, economy, public health, and security (Guo et al., 2019).

In this context, automatic misinformation detection has attracted the interest of different communities. In a broad definition, misinformation detection (MID) is the task of assessing the appropriateness (truthfulness, credibility, veracity or authenticity) of claims in a piece of information (Su et al.,

Cabral, L., Monteiro, J., Franco da Silva, J., Mattos, C. and Mourão, P.
FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages.
DOI: 10.5220/0010446800630074
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 63-74
ISBN: 978-989-758-509-8

2020). Early detection of misinformation could prevent it's spread, thus reducing its damage. MID can be exploited by various approaches, including human-crafted rules, traditional machine learning models, neural networks and combining machine learning and natural language processing (NLP).

The combination of machine learning and natural language processing (NLP) to extract features from text achieved great results in the literature. NLP-based approaches rely on the hypothesis that intentionally misleading texts has linguistic patterns that distinguishes from non-misleading texts. This approach has been extensively used with data collected from platforms as Facebook[1] (Granik and Mesyura, 2017) and Twitter[2] (Zervopoulos et al., 2020). However, in many developing countries such as Brazil, India, and Mexico, one of the primary sources of misinformation is the messaging application WhatsApp[3]. The purpose of WhatsApp is allow users to privately send messages to each other through their smartphones. Despite being mostly used for individual conversations, WhatsApp has the resources of conversation groups, where up to 256 users can participate, and forwarding messages, which facilitate the quick dissemination of misinformation. In Brazil's case, about 35% deceptive news is shared through WhatsApp (Newman et al., 2020), and 40,7% of these messages are shared after being disproved (Resende et al., 2018; Resende et al., 2019).

Despite this scenario, due to the private messaging nature of WhatsApp, there still few methods of MID developed specifically for this platform. When comes to NLP-based approaches, the performance of a model is highly dependent on the linguistic patterns, topics, and vocabulary present in the data used to train it. Due to its unique nature of private messenger and its broad user base, the content shared through WhatsApp and the way its users express themselves varies significantly compared to public social networks like Facebook and Twitter (Waterloo et al., 2018; Rosenfeld et al., 2018). Then, a model trained with texts collected from Twitter or Facebook may have a poor performance when used to classify WhatsApp messages. Thus, in this context, to obtain a good NLP-based MID is necessary to train the prediction model with WhatsApp data.

In order to fill this gap, we built a large-scale, labeled, anonymized, and public dataset formed by WhatsApp messages in Brazilian Portuguese (PT-BR), collected from public WhatsApp groups. Then, we conduct a series of classification experiments using combinations of Bag-Of-Words features and classical machine learning methods to answer the following research questions:

1. How challenging is the task of misinformation detection in WhatsApp messages using NLP-based techniques?
2. Which combination of pre-processing methods, word-level features and classification algorithms are best suited for this task?
3. Which are the limitations of an NLP-based approach?

Our results show that a purely NLP-based approach using traditional Bag-of-Words features has limited performance due to the particularities of WhatsApp messages, especially the predominance of short messages that follows media files (audios, images, or videos). Our best result achieved a F1 score of 0.73, and the analysis of errors indicates that they occur mainly due to the predominance of short texts that accompany media files. When texts with less than 50 words are filtered, the F1 score rises to 0.87. To the best of our knowledge, there is no previous work that performed MID in a large-scale corpus of WhatsApp messages in PT-BR.

The remainder of this paper is organized as follows. Section 2 presents the main related work. Section 3 describes the process used to create a large-scale, labeled, anonymized, and public dataset of WhatsApp messages in PT-BR. Section 4 details our experimental setup for MID. Section 5 reports and discuss the results. Conclusions and future work are presented in 6.

2 RELATED WORK

Several works attempt to detect misinformation in different languages and platforms. Most of them use news in English or Chinese languages. Besides, Websites and social media platforms with easy access to data, like Twitter, for example, are amongst the main sources used to build misinformation datasets.

Despite the large number of works investigating the misinformation detection problem, few of the search for suitable solutions for the Brazilian Portuguese language (PT-BR). In this context, (Monteiro et al., 2018) presented the first and largest Fake News' corpus in Brazilian Portuguese (PT-BR), called Fake.Br. This corpus was built manually, collecting Fake News on the Web and, semi-automatically, searching for the actual news related to each Fake News, generating an equal amount of negative and positive examples. In all, the dataset has 7,200 items

[1] https://www.facebook.com/

[2] https://twitter.com/

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

(or news), with 3,600 true and 3,600 false. In addition, the authors evaluated some classifiers (Naıve-Bayes, Random Forest, and Multilayer Perceptron). After, the work presented in (Silva et al., 2020) investigated the use of different features and algorithms in order to detect fake news, exploring the Fake.Br corpus (Monteiro et al., 2018).

However, it is important to highlight that WhatsApp is unique in several ways relative to other social networks. A particularly novel aspect of WhatsApp messaging is its close integration with large public groups. These are openly accessible groups, frequently publicized on well-known websites, and typically themed around particular topics, like politics, religion, soccer, etc. The study presented in (Garimella and Tyson, 2018) is a pioneering work in collecting and analyzing WhatsApp' messages. The authors built a dataset by crawling 178 public groups, containing 45K users and 454K messages, from different countries and languages, such as India, Pakistan, Russia, Brazil, and Colombia. Nevertheless, no solution to the misinformation detection problem was presented. In (Gaglani et al., 2020), the authors contextualize the problem of spreading fake news on WhatsApp, especially in India and Brazil, and proposes a strategy for the automatic detection of fake news. A total of 10 group chats are scraped for one week to get 1000 multilingual messages. After cleaning the data, the multilingual data was translated into English by employing the google translate API. So, the proposed approach for misinformation detection does not consider the particularities of each language.

Thus, despite the efforts of the scientific community, there is still a need for a large-scale corpus containing WhatsApp messages in Portuguese. It is worth mentioning that texts extracted from WhatsApp are quite different from those collected through Websites, fact-checkers, or other kinds of social media platforms, such as Twitter. WhatsApp messages include conversation, opinions, humorous and satirical texts, prayers, commercial offers, news, short texts, emojis, and others. Then, using the Fake.Br corpus, for example, to automatic misinformation detection in WhatsApp is not a suitable approach. In this scenario, (Faustini and Covões, 2019) is a seminal work. In the experiments, three different datasets were explored in order to detect fake news: Fake.Br (news from websites), a Twitter corpus, obtained using the Twitter API, and a small WhatsApp corpus. It is worth mentioning that the WhatsApp corpus was obtained from texts on the website boatos.org and have only 177 messages, where 165 are fake and 12 are true. Some papers present a few initiatives in order to gathering, analyzing, and visualize public groups in WhatsApp (Resende et al., 2018; Machado et al., 2019; Resende et al., 2019). Nevertheless, the collected data were not labeled, no dataset has been made publicly available, and no solution to the misinformation detection problem was presented. Table 1 shows a comparative analysis between the datasets of WhatsApp messages in Brazilian Portuguese found in the literature.

3 THE FakeWhatsApp.Br DATASET

In order to develop automatic misinformation detection approaches, that are suitable for WhatsApp messages in Brazilian Portuguese, a critical aspect is a need for a large-scale labeled dataset. However, to the best of our knowledge, there is no corpus for Brazilian Portuguese with these characteristics. To fill the gap of the lack of a large-scale labeled corpus of WhatsApp messages in Brazilian Portuguese we built the FakeWhatsApp.Br, inspired by (Silva et al., 2020).

The work of (Rubin et al., 2015) suggests a methodological guideline for building corpora of deceptive content, which includes: the corpus must contain truthful texts and their corresponding untruthful versions, in order to allow finding patterns and regularities in “positive and negative instances”; the texts in the corpus should be in plain text format; the texts should have similar sizes to avoid bias in learning; the texts should belong to a specific time interval, as writing style changes in time; and the corpus should keep the related metadata information (e.g., the URL of the news, the authors, publication date, and number of comments and visualizations) because it can be useful for fact checking algorithms.

3.1 Data Collecting

Unlike other social media, such as Twitter and Facebook, and due to its private chat nature, there is no public API to collect data from WhatsApp in an automated manner. Thus, creating a dataset of WhatsApp messages poses a technical, and even ethical, challenge. To tackle this issue, we take an approach similar to (Garimella and Tyson, 2018; Resende et al., 2018).

Initially, we seek for public groups with political themes during the Brazilian general elections campaign in 2018. The groups were found by searching for “chat.whatsapp.com/” on the Web and manually analyzing its content. We established a rule to join only groups with at least 100 or more users to explore only relevant content groups, whereas WhatsApp has a limitation of a maximum of 256 users by

Table 1: Datasets of WhatsApp Messages in Brazilian Portuguese. Hyphen (-) means that the information could not be found in the work.

Work	Labeled	Total of Text Messages	Groups	Users	MID	Publicly Available
(Faustini and Covões, 2019)	Yes	177	-	-	Yes	Yes
(Resende et al., 2018)	No	169,154	127	6,314	No	No
(Machado et al., 2019)	No	298,892	130	-	No	No
(Resende et al., 2019)/ Truck Drivers' Strike	No	95,424	141	5,272	No	No
(Resende et al., 2019)/ Election Campaign	No	591,162	136	18,725	No	No
FakeWhastApp.BR	Yes	5,284	59	14,784	Yes	Yes

group. After careful selection, we joined 59 public groups. Next, we created a WhatsApp account to join the selected groups. So, we collected messages from July to November of 2018. After this period, we extracted all content and metadata, building a data matrix, where each row corresponds to a message sent in a group. The matrix columns are the date and hour that the message was sent, the sender's phone number, the international phone code, the Brazilian state (if the user is from Brazil), the content (text) of the message, the word and character counts, and if the message contained media such as audio, image or video. Nevertheless, since we are finding to identify misinformation in the WhatsApp message text, the FakeWhatsApp.Br dataset does not contain media files. Besides, we also count how many times the same message text appears in the dataset. For doing so, we only consider messages with identical textual content that had more than five words, to filter common messages such as greetings. We call the messages in which the textual content appears more than once in the dataset "viral messages".

3.2 Data Anonymization

We took into consideration privacy issues by anonymizing users' names and cell phone numbers. For this, we create an anonymous and unique ID for each user by using a hash function on their phone number. Similarly, we create an anonymous alias for each group. Since the groups are public, our approach does not violate WhatsApp's privacy policy [4].

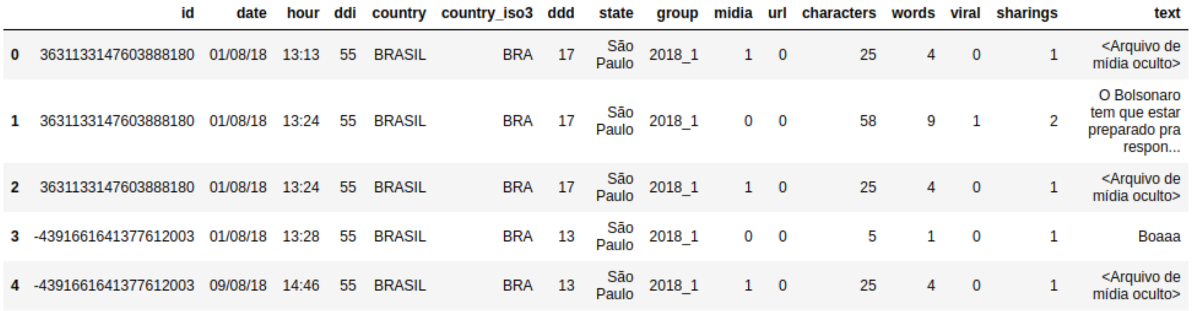
Figure 1 illustrates the FakeWhatsApp.Br dataset at this time, before data labeling. The FakeWhatsApp.Br dataset has 282,601 WhatsApp messages from users and groups from all Brazilian states. It is important to note that although FakeWhatsApp.Br dataset has several metadata associated with each message, in this work we will use exclusively the textual data in order to build misinformation detection models. However, these metadata will be used in future works to improve the performance of the proposed misinformation detection approaches.

3.3 Corpus Labeling

Building a large-scale dataset is one of the biggest challenges for the automatic detection of misinformation. However, data labeling is another challenge because we need to specify whether a part of the text is true or false based on the truth. Notes can generally be made by specialized journalists or fact-checking sites.

Next, we will describe the used method for labeling the WhatsApp messages' textual content. In order to create a high-quality corpus, the process used for data labeling was entirely manual. A human specialist checked the content of each message and determined if it contains misinformation or not. Since this process is time-consuming, we chose to labeled only the unique viral messages, resulting in a much smaller subset with 5,284 unique messages. This decision is backed by the work of (Vosoughi et al., 2018), where is shown that misinformation spreads faster, deeper, and wider in social networks than true information. We argue that in that way we avoid having peer-to-peer conversation data in the corpus, allowing us to create and validate classification models focused on detecting misinformation which are most spread and harmful. The subset of viral messages contains a va-

[4] https://www.whatsapp.com/legal/privacy-policy



	id	date	hour	ddi	country	country_iso3	ddd	state	group	midia	url	characters	words	viral	sharings	text
0	3631133147603888180	01/08/18	13:13	55	BRASIL	BRA	17	São Paulo	2018_1	1	0	25	4	0	1	<Arquivo de mídia oculto>
1	3631133147603888180	01/08/18	13:24	55	BRASIL	BRA	17	São Paulo	2018_1	0	0	58	9	1	2	O Bolsonaro tem que estar preparado pra respon...
2	3631133147603888180	01/08/18	13:24	55	BRASIL	BRA	17	São Paulo	2018_1	1	0	25	4	0	1	<Arquivo de mídia oculto>
3	-4391661641377612003	01/08/18	13:28	55	BRASIL	BRA	13	São Paulo	2018_1	0	0	5	1	0	1	Boaaa
4	-4391661641377612003	09/08/18	14:46	55	BRASIL	BRA	13	São Paulo	2018_1	1	0	25	4	0	1	<Arquivo de mídia oculto>

Figure 1: Sample from the collected data before labelling.

riety of types of messages, such as fake news, rumors, hoaxes, true news, political advertising, opinions, satires, jokes, election polls and hate speech. We classified all theses messages with the general definition of misinformation adopted in (Su et al., 2020) and consider the labels 1 (contains misinformation) and 0 (does not contain misinformation).

As follows we summarize our labeled guideline and offer parts of messages as examples (and it's translations) for each situation, without the emoticons:

1. If the text contains verifiable untrue claims, we annotate it as misinformation. For that purpose, we made extensive use of trustful Brazilian's fact-checking platforms, such as *Agência Lupa*[5] and *Boatos.org*[6].

   E.g.: "Bolsa Ditadura se transformou em indústria: VC sabia que 20mil anistiados, entre eles, Chico Buarque, Gilberto Gil, Caetano Veloso, Marieta Severo, Taiguara, Lula, Zé Dirceu, Fernando Henrique Cardoso, recebem o Bolsa Ditadura mensalmente e são isentos de pagar Imposto de Renda? Sendo que dos 20 mil, 10 mil recebem indenizações mensais acima do teto constitucional(R$ 33.763,00) Essa esquerda maldita tira dos cofres públicos mensalmente a bagatela de R$ 365.000.000,00 (Trezentos e sessenta e cinco milhões) pagos por nós, otários!"[7].

   Translation: "Dictatorship Grant turned into industry: did you know that 20,000 amnesties, including Chico Buarque, Gilberto Gil, Caetano Veloso, Marieta Severo, Taiguara, Lula, Zé Dirceu, Fernando Henrique Cardoso, receive the Dictatorship Grant monthly and are exempt from paying Tax Income? Of the 20 thousand, 10 thousand receive monthly indemnities above the constitutional ceiling (R$ 33,763.00) This cursed left removes from the public coffers monthly the trifle of R$ 365,000,000.00 (Three hundred and sixty-five million) paid for us suckers!"

2. If the text contains claims that cannot be proven and that are imprecise, biased, alarmist or are harmful to groups or individuals, we annotate it as misinformation.

   E.g.: "O golpe da esquerda é o seguinte: a viagem do Ciro Gomes à Europa foi proposital! Um teatro p/ colocar a seguinte narrativa em prática: ele sai de cena, ou seja, teoricamente não está apoiando Haddad, de repente ele volta (e de fato, de acordo c/ o Estadão ele está chegando hoje) qdo não terá mais nenhuma pesquisa a ser divulgada, para q não se "comprove", se de fato aconteceria, a escalada q Haddad "terá" de milhões de votos em 2 dias. Enfim, o fato novo, que a imprensa já avisadamente, a todo tempo publicou, que seria a única coisa p/ virada nos votos! Ou seja, tudo articulado, para acontecer exatamente como no 1° turno, onde o apoio do Lula fez o poste crescer em poucos dias 20 ptos percentuais, vão tentar vender a ideia que o apoio do Ciro de última hora, fez reverter ao Haddad todos os votos que ele teve, justificando o golpe nas urnas! Precisamos divulgar isto em massa, numa velocidade recorde, para minar o efeito, antes mesmo de ocorrer! (...)"

   Translation: "The coup of the left is as follows: Ciro Gomes' trip to Europe was purposeful! A act to put the following narrative into practice: he leaves the scene, that is, theoretically he is not supporting Haddad, suddenly he comes back (and in fact, according to Estadão he is arriving today) when he will have no more election pools to be released, so that it is not "proven", if in fact it would happen, the escalation that Haddad "will have" of millions of votes in 2 days. Anyway, the new fact, which the press has already warned, published all the time, which would be the only thing to change the votes! That is, everything articulated, to hap-

[5] http://piaui.folha.uol.com.br/lupa/

[6] http://www.boatos.org/

[7] https://www.aosfatos.org/noticias/nao-e-verdade-que-governo-paga-bolsa-ditadura-20-mil-anistiados-politicos/

pen exactly as in the 1st round, where Lula's support made the post grow in a few days by 20 percentage points, they will try to sell the idea that Ciro's last-minute support, made Haddad revert all the votes he had, justifying the coup in the elections! We need to disclose this en masse, at record speed, to undermine the effect, even before it occurs! (...)".

3. By decision of the Brazilian Superior Electoral Court, informal electoral polls, which do not meet formal requirements and scientific rigors, were banned in the 2018 elections. Thus, we annotate messages containing such polls as misinformation.

   E.g.: "Vota aí e repassa!!! Vamos ver se o ibope está certo? https://pt.surveymonkey.com/r/W85R38F"

   Translation: "Vote and forward!!! Let's see if the IBOPE is right? https://pt.surveymonkey.com/r/W85R38F"

4. Some of the messages are short texts originally accompanied by media content (image, audio, or video) which not readily accessible. In those cases, we search on the Web for the media content and, if we find the media, we assign a label following the previous criteria.

   E.g.: "Antes de decidir seu voto ouça o que diz o padre Marcelo Rossi". [8]

   Translation: "Before deciding your vote, listen to what Father Marcelo Rossi says".

5. If the original media content cannot be found, we look for indications of Item 2 in the text itself.

   E.g.: "Olha o que os partidos de esquerda defendem E se votarmos viraremos isso".

   Translation: "Look what the leftist parties defend. And if we vote we will turn it into this".

6. If none of the previous indications is found in the text, we consider it as not containing misinformation. We take careful consideration when the text is an opinion instead of a claim or is humorous, assigning a non-misinformation label in both cases.

   E.g.: "Relaxando no sofá, barriguinha plusize, 9mm na cintura, sem coldre, no pelo, com saque cruzado, relogio Cassio modelo 1985 no punho e xingando comunistas no insta... Esse é meu Presidente!".

   Translation: "Relaxing on the couch, plus size tummy, 9mm at the waist, no holster, with crossed loot, Cassio model 1985 watch on the wrist and cursing communists at the Instagram... This is my President!".

During the labeled process, we observed that some of the messages text could be found on other social media, such as YouTube, Twitter, and Facebook. Out of a total of 5,284 messages, 610 (11.5%) could be found on different media. Out of these, 85 (14%) were found on Twitter, 236 (38.7%) on Facebook, 240 (39.3%) on YouTube. The remaining 49 (8%) were found on various Web pages, like blogs, news portals, etc. The majority of the messages were exclusive to WhatsApp. Some of these use a formatting specific to the platform, e.g., the underscore on both sides of the text used to format the text as italic, or the asterisk on both sides of the text to bold it. A high quantity and variety of emoticons were also perceived in some messages, thus reinforcing the evidence that WhatsApp messages have their particularities.

After the labeling process, the FakeWhatsApp.BR corpus contains 2,193 unique messages annotated as misinformation (label 1) and 3,091 unique messages annotated as non-misinformation (label 0). In Table 2, we present basic statistics about the corpus, including some traditional NLP features based on the number of tokens, types, characters, as well as the average number of shares, i.e., the frequency of the message in the original dataset.

As expected, the messages labeled as misinformation were, on average, more shared in the groups. We can see in Table 2 that the majority of messages of the corpus are short texts, but the distribution of the number of tokens have a heavy tail, with a minority of very long texts. We also point out that the average number of tokens and types is much higher in the messages with misinformation. This difference in the size of the messages can be problematic for machine learning classification algorithms, creating a bias about the text size (Rubin et al., 2015).

4 EXPERIMENTAL EVALUATION

To answer the research questions presented in Section 1 and provide a baseline for the misinformation detection problem in WhatsApp messages in Brazilian Portuguese, we carefully designed a set of experiments using the FakeWhatsApp.Br dataset. We have explored different combinations between features and classification algorithms. To obtain robust statistical results, we performed our experiments using k-fold cross-validation, with $k = 5$ folds.

[8] https://www.boatos.org/religiao/padre-marcelo-rossi-grava-audio-brasil-bolsonaro-comunismo.html

Table 2: FakeWhatsApp.Br basic statistics.

Statistics	Non-misinformation	Misinformation
Count of unique messages	3,091	2,193
Mean and standard deviation of number of tokens in messages	51.27 ± 126.28	106.55 ± 169.31
Minimum number of tokens	6	6
Median number of tokens	20	34
Maximum number of tokens	2,664	2,203
Mean and standard deviation of number of types in messages	38.03 ± 66.44	73.78 ± 97.79
Average size of words (in characters)	6.24	5.66
Type-token ratio	0.91	0.85
Mean and standard deviation of shares	3.32 ± 3.90	4.83 ± 6.81

4.1 Features and Classification Algorithms

Different text's feature extraction methods were evaluated. However, we focus our experiments in traditional Bag-Of-Words (BoW) text representation. We choose to not use pre-trained embedding vectors due to the large presence of misspelled words, emoticons and neologisms in the corpus, thus resulting in several out-of-vocabulary words. In addition, we seek to establish a baseline for the automatic misinformation detection problem in WhatsApp messages in Brazilian Portuguese. So, BoW features are suitable for this purpose due to its simplicity, processing speed and its wide use in text classification problems.

Then, we explored vectors created with binary BoW and with TF-IDF, converting the text to lowercase and using whitespaces and punctuation marks as token separators. Emojis are abundant in the texts and are part important of the dialect used in WhatsApp and so we chose to keep them in tokenization. However, as combinations of emojis can generate different kinds of tokens, we separate all emojis with whitespaces, thus creating a token for each emoji. We also normalize URLs, maintaining only it's domains name. In the same manner, we normalize the Brazilian's text laugh, written as a sequence with a varying number of the letter k, which we convert to a unique dummy feature "kkkk" (somewhat equivalent to the english's "LoL"). Due to the corpus' lexical diversity, the resulting vectors have large dimension and sparsity.

Still, besides using only unigrams as tokens, we also varied the n-gram range, experimenting the combination of unigrams, bigrams and trigrams. Even if this results in a larger vector space, from our knowledge of the domain, we believe that the combination of bigrams and trigrams can reveal distinguishable patterns that are present in messages with misinformation in our dataset. Lastly, to compare the impact of more advanced pre-processing techniques to reduce vector space, we include a set of experiments with utilizes steps of lemmatization and stop words removal in the pre-processing.

Thus, we combine these different vectorization approaches (binary BoW or TF-IDF), the n-grams range (unigrams, bigrams and trigrams) and the use of extra steps of pre-processing (lemmatization and stop words removal), creating a total of 12 different features scenarios.

In each of these scenarios, we perform experiments with a selection of 9 machine learning classification algorithms, broadly used in text classification tasks (Pranckevičius and Marcinkevičius, 2017): logistic regression (LR), Bernoulli (if the features are BoW) or Complement Naive-Bayes (if features are TF-IDF) (NB) (Kim et al., 2006; Rennie et al., 2003), support vector machines with a linear kernel (LSVM), SVM trained with stochastic gradient descent (SGD), SVM trained with a RBF kernel (Prasetijo et al., 2017) (SVM), K-nearest neighbors (KNN), random forest (RF), gradient boosting (GB) and multilayer perceptron neural network (MLP).

For all algorithms, we used the implementation from the Python library scikit-learn (Pedregosa et al., 2011). The MLP used a batch size of 64 and a early stopping training strategy, where 10% of training data is set aside as validation and terminate training when validation score is not improving by at least 0.001 for 5 consecutive epochs. All the others hyperparameters for this and the others models are used as the default. It is important to note that the chosen set of algorithms encompasses different families of machine learning algorithms: linear models (LR), generative models (NB), instance-based learning (KNN), support vector machines (LSVM, SVM and SGD), ensemble methods - bagging (RF) and boosting (GB), and neural networks (MLP). Although we do not perform a systematic selection of hyperparameters for each model, the variety of the tested approaches should give us information of which learning strategy can be more well suited to this problem and establishes a baseline.

Considering all combinations between features, pre-processing and classification methods, we performed a total of 108 experiments, which should give us information to answer Research Question 1 and Research Question 2.

4.2 Performance Metrics

To evaluate the performance of each experiment, we adapt the metrics used in (Silva et al., 2020) considering the formulation of our problem and goal. As mentioned previously, we tackle the problem as a binary classification task, where the misinformation represents the positive class (and also the class of interest) and the non-misinformation represents the negative class. Below we list the chosen evaluation metrics:

- False Positive Rate (FPR): the proportion of messages without misinformation incorrectly classified as misinformation. The lower, the better.
- Precision (PRE): proportion of messages classified as misinformation and that truly belong to the misinformation class. The higher, the better.
- Recall (REC): proportion of misinformation correctly classified. The higher, the better.
- F1-score (F1): harmonic average between precision and recall.

As we use a k-fold cross validation, the mean and standard deviation of each metric will be presented. After these experiments, we choose the best classifier and features, retrain it with a randomly separated train set (80% of the total data) and test it on the remaining 20% of the data. To answer Research Question 3, we did a qualitative analysis of the false positive and false negative results of the best classifier, identifying and categorizing the possible reasons of the errors, and so the limitations of a NLP-based approach.

5 RESULTS

In order to allow future research work in this task, as well for reproducibility of the experiments, the source code and the FakeWhatsApp.Br corpus are publicly available at a public online repository[9].

The experimental results are summarized in Tables 3 and 4, where we present the results for BoW and TF-IDF features, respectively. In each Table we present the results of each features' scenario that vary with the n-gram range and with the use of lemmatization and stopwords removal. The sub tables in each Table are organized as follows:

a) only unigrams;

b) unigrams and bigrams;

c) unigrams, bigrams and trigrams;

d) only unigrams, stopwords removal and lemmatization;

e) unigrams and bigrams, stopwords removal and lemmatization;

f) unigrams, bigrams and trigrams, stopwords removal and lemmatization;

From Tables 3 and 4 we can note that none classifier was always superior in every scenarios. However, the MLP, LSVM, SGD and LR methods performed consistently well in all scenarios. In the other hand, the NB, KNN and RF methods had the worst results, considering the F1-score.

Although the difference between the best scores is low (3.2% improvement from the maximum to the minimum), we can see that the results did improve with the use of bigrams and trigrams. Comparing the scores in sub tables **a)**, **b)** and **c)**, from both Tables, we see consistent improvement of the results. As we expected, bigrams and trigrams tokens contains relevant information in this domain, relative to frequent topics in messages with misinformation during the time-period in which the data was collected.

Similarly, when we compare sub tables **a)**, **b)**, **c)** with **e)** and **f)**, we see that the use of lemmatization and stop words removal also slightly improved the scores only when using bigrams and trigrams. As for the vectorization method, BoW features had a better performance when using only unigrams (sub tables**a)** and **d)**) and were outperformed by TF-IDF features when using bigrams and trigrams.

Table 5 summarizes the top 10 best results for all the experiments. Considering the F1-score the results are very close, and allows us to see that the best results were achieved with TF-IDF, a higher n-gram range, the removal of stopwords and lemmatization, as well as the use of LSVM, MLP and SGD methods.

From the results, we can assess how challenging the problem of detecting misinformation in WhatsApp is, answering the Research Question 1, since we did not obtain any F1 score above 0.74. Comparing with the results obtained by (Silva et al., 2020) in the Fake.Br corpus, which obtained a F1 score of 0.965 using a combination of TF-IDF and linguistic features with an ensemble strategy, we see room for improvement. However, it's important to highlight that even the problems are similar, the datasets contains many

[9] https://github.com/cabrau/FakeWhatsApp.Br.

Table 3: Results with binary BoW features. MLP, LSVM, SGD and LR methods performed consistently well in most scenarios. In general, the results improved when stopword removal, lemmatization and bigrams and trigrams were used.

a) BOW-1. Features: 23,422

Model	FPR	PRE	REC	F1
LR	**0.183 ± 0.00**	0.761 ± 0.02	0.677 ± 0.02	**0.715 ± 0.00**
NB	0.229 ± 0.00	0.753 ± 0.01	0.342 ± 0.03	0.469 ± 0.03
LSVM	0.205 ± 0.01	0.718 ± 0.01	0.674 ± 0.02	0.695 ± 0.01
SGD	0.202 ± 0.01	0.718 ± 0.02	**0.704 ± 0.02**	0.710 ± 0.01
SVM	0.185 ± 0.01	0.786 ± 0.01	0.593 ± 0.06	0.674 ± 0.04
KNN	0.219 ± 0.00	**0.850 ± 0.04**	0.291 ± 0.02	0.433 ± 0.02
RF	0.189 ± 0.01	0.813 ± 0.02	0.513 ± 0.06	0.626 ± 0.05
GB	0.197 ± 0.00	0.765 ± 0.01	0.565 ± 0.03	0.649 ± 0.01
MLP	0.192 ± 0.01	0.754 ± 0.03	0.652 ± 0.05	0.696 ± 0.01

b) BOW-1,2. Features: 156,182

Model	FPR	PRE	REC	F1
LR	**0.180 ± 0.01**	0.775 ± 0.02	0.655 ± 0.03	**0.709 ± 0.01**
NB	0.222 ± 0.00	0.834 ± 0.02	0.281 ± 0.02	0.420 ± 0.03
LSVM	0.188 ± 0.01	0.757 ± 0.03	0.657 ± 0.02	0.702 ± 0.01
SGD	0.198 ± 0.01	0.728 ± 0.01	**0.683 ± 0.02**	0.705 ± 0.02
SVM	0.186 ± 0.01	0.819 ± 0.01	0.520 ± 0.07	0.633 ± 0.05
KNN	0.220 ± 0.00	**0.871 ± 0.04**	0.267 ± 0.04	0.406 ± 0.05
RF	0.193 ± 0.01	0.826 ± 0.01	0.469 ± 0.06	0.595 ± 0.04
GB	0.202 ± 0.00	0.758 ± 0.01	0.551 ± 0.04	0.636 ± 0.02
MLP	0.184 ± 0.00	0.765 ± 0.03	0.664 ± 0.05	0.708 ± 0.02

c) BoW-1,2,3. Features: 384,783

Model	FPR	PRE	REC	F1
LR	**0.179 ± 0.01**	0.782 ± 0.02	0.641 ± 0.03	0.704 ± 0.01
NB	0.256 ± 0.01	0.637 ± 0.01	**0.804 ± 0.02**	0.710 ± 0.01
LSVM	0.185 ± 0.01	0.772 ± 0.03	0.633 ± 0.03	0.695 ± 0.01
SGD	0.197 ± 0.01	0.725 ± 0.02	0.710 ± 0.04	**0.716 ± 0.02**
SVM	0.193 ± 0.01	0.844 ± 0.01	0.445 ± 0.06	0.579 ± 0.05
KNN	0.220 ± 0.00	**0.890 ± 0.04**	0.255 ± 0.04	0.394 ± 0.05
RF	0.194 ± 0.00	0.850 ± 0.02	0.432 ± 0.05	0.570 ± 0.04
GB	0.202 ± 0.00	0.757 ± 0.01	0.551 ± 0.04	0.637 ± 0.02
MLP	0.188 ± 0.01	0.762 ± 0.03	0.645 ± 0.05	0.696 ± 0.02

d) BOW-1-STOPWORDS-LEMMA. Features: 19,455

Model	FPR	PRE	REC	F1
LR	**0.181 ± 0.00**	0.764 ± 0.01	0.678 ± 0.02	0.718 ± 0.00
NB	0.228 ± 0.00	0.764 ± 0.02	0.327 ± 0.03	0.457 ± 0.02
LSVM	0.203 ± 0.00	0.716 ± 0.01	**0.690 ± 0.02**	0.703 ± 0.01
SGD	0.198 ± 0.01	0.731 ± 0.01	0.669 ± 0.01	0.699 ± 0.01
SVM	0.186 ± 0.01	0.784 ± 0.01	0.588 ± 0.06	0.670 ± 0.03
KNN	0.219 ± 0.00	**0.829 ± 0.03**	0.311 ± 0.03	0.451 ± 0.03
RF	0.182 ± 0.00	0.803 ± 0.01	0.576 ± 0.05	0.669 ± 0.03
GB	0.196 ± 0.00	0.783 ± 0.00	0.529 ± 0.05	0.630 ± 0.03
MLP	0.182 ± 0.00	0.760 ± 0.01	0.684 ± 0.02	**0.720 ± 0.00**

e) BOW-1,2-STOPWORDS-LEMMA. Features: 129,745

Model	FPR	PRE	REC	F1
LR	**0.173 ± 0.00**	0.794 ± 0.01	0.648 ± 0.02	0.713 ± 0.01
NB	0.224 ± 0.00	**0.879 ± 0.03**	0.238 ± 0.01	0.374 ± 0.01
LSVM	0.180 ± 0.00	0.774 ± 0.01	0.655 ± 0.02	0.709 ± 0.01
SGD	0.184 ± 0.01	0.764 ± 0.03	0.663 ± 0.05	0.708 ± 0.02
SVM	0.186 ± 0.00	0.833 ± 0.01	0.497 ± 0.05	0.621 ± 0.04
KNN	0.217 ± 0.00	0.858 ± 0.04	0.297 ± 0.04	0.438 ± 0.04
RF	0.189 ± 0.01	0.848 ± 0.01	0.461 ± 0.05	0.595 ± 0.04
GB	0.198 ± 0.00	0.778 ± 0.00	0.522 ± 0.03	0.624 ± 0.02
MLP	**0.173 ± 0.00**	0.781 ± 0.00	**0.678 ± 0.02**	**0.726 ± 0.01**

f) BoW-1,2,3-STOPWORDS-LEMMA. Features: 261,665

Model	FPR	PRE	REC	F1
LR	0.172 ± 0.00	0.806 ± 0.02	0.625 ± 0.03	0.703 ± 0.02
NB	0.273 ± 0.01	0.620 ± 0.01	**0.819 ± 0.01**	0.706 ± 0.01
LSVM	0.176 ± 0.01	0.790 ± 0.02	0.640 ± 0.02	0.707 ± 0.02
SGD	0.184 ± 0.01	0.770 ± 0.02	0.642 ± 0.05	0.698 ± 0.02
SVM	0.195 ± 0.00	0.861 ± 0.01	0.415 ± 0.05	0.558 ± 0.04
KNN	0.217 ± 0.00	**0.902 ± 0.04**	0.259 ± 0.04	0.400 ± 0.05
RF	0.195 ± 0.01	0.853 ± 0.02	0.425 ± 0.05	0.565 ± 0.04
GB	0.197 ± 0.00	0.783 ± 0.00	0.521 ± 0.03	0.625 ± 0.02
MLP	**0.171 ± 0.00**	0.795 ± 0.02	0.657 ± 0.03	**0.718 ± 0.01**

Table 4: Results with TF-IDF features. In general, the results improved slightly when compared to binary BoW features, especially when using bigrams and trigrams, stopword removal and lemmatization. The same classification methods stood out: MLP, LSVM, SGD.

a) TFIDF-1. Features: 23,422

Model	FPR	PRE	REC	F1
LR	0.197 ± 0.01	0.745 ± 0.03	0.636 ± 0.04	0.685 ± 0.02
NB	0.229 ± 0.00	0.753 ± 0.01	0.342 ± 0.03	0.469 ± 0.03
LSVM	0.197 ± 0.00	0.726 ± 0.01	0.703 ± 0.02	**0.714 ± 0.01**
SGD	0.203 ± 0.01	0.715 ± 0.02	0.705 ± 0.02	0.710 ± 0.01
SVM	**0.184 ± 0.00**	0.773 ± 0.02	0.635 ± 0.04	0.696 ± 0.02
KNN	0.323 ± 0.01	0.569 ± 0.01	**0.727 ± 0.01**	0.638 ± 0.01
RF	0.196 ± 0.00	**0.809 ± 0.03**	0.485 ± 0.04	0.605 ± 0.02
GB	0.203 ± 0.00	0.746 ± 0.02	0.581 ± 0.03	0.652 ± 0.01
MLP	0.201 ± 0.01	0.720 ± 0.02	0.701 ± 0.00	0.710 ± 0.01

b) TFIDF-1,2. Features: 156,182

Model	FPR	PRE	REC	F1
LR	0.206 ± 0.01	0.725 ± 0.02	0.633 ± 0.03	0.675 ± 0.01
NB	0.222 ± 0.00	0.834 ± 0.02	0.281 ± 0.02	0.420 ± 0.03
LSVM	0.197 ± 0.01	0.721 ± 0.02	0.729 ± 0.03	0.724 ± 0.02
SGD	0.203 ± 0.01	0.709 ± 0.02	**0.731 ± 0.03**	0.720 ± 0.01
SVM	0.195 ± 0.01	0.752 ± 0.02	0.623 ± 0.05	0.680 ± 0.02
KNN	0.313 ± 0.02	0.577 ± 0.02	0.715 ± 0.03	0.639 ± 0.02
RF	**0.192 ± 0.00**	**0.835 ± 0.02**	0.462 ± 0.04	0.593 ± 0.03
GB	0.201 ± 0.00	0.752 ± 0.02	0.578 ± 0.03	0.652 ± 0.01
MLP	0.197 ± 0.01	0.722 ± 0.03	0.730 ± 0.03	**0.725 ± 0.01**

c) TFIDF-1,2,3. Features: 384,783

Model	FPR	PRE	REC	F1
LR	0.216 ± 0.01	0.699 ± 0.03	0.676 ± 0.03	0.686 ± 0.02
NB	0.245 ± 0.01	0.651 ± 0.01	0.753 ± 0.03	0.697 ± 0.01
LSVM	0.210 ± 0.01	0.695 ± 0.02	0.767 ± 0.03	0.729 ± 0.01
SGD	0.212 ± 0.02	0.691 ± 0.02	**0.776 ± 0.02**	**0.731 ± 0.02**
SVM	0.203 ± 0.01	0.729 ± 0.02	0.648 ± 0.04	0.685 ± 0.02
KNN	0.309 ± 0.02	0.581 ± 0.02	0.722 ± 0.02	0.644 ± 0.02
RF	**0.198 ± 0.00**	**0.829 ± 0.02**	0.437 ± 0.04	0.570 ± 0.03
GB	0.207 ± 0.00	0.737 ± 0.03	0.588 ± 0.04	0.651 ± 0.01
MLP	0.212 ± 0.01	0.695 ± 0.02	0.751 ± 0.03	0.721 ± 0.01

d) TFIDF-1-STOPWORDS-LEMMA. Features: 19,455

Model	FPR	PRE	REC	F1
LR	0.188 ± 0.00	0.766 ± 0.01	0.621 ± 0.03	0.685 ± 0.01
NB	0.228 ± 0.00	0.764 ± 0.02	0.327 ± 0.03	0.457 ± 0.02
LSVM	0.196 ± 0.00	0.729 ± 0.01	**0.694 ± 0.02**	**0.711 ± 0.00**
SGD	0.199 ± 0.00	0.723 ± 0.01	0.694 ± 0.01	0.708 ± 0.00
SVM	**0.181 ± 0.00**	0.785 ± 0.01	0.620 ± 0.02	0.692 ± 0.01
KNN	0.241 ± 0.01	0.660 ± 0.02	0.643 ± 0.03	0.651 ± 0.02
RF	0.183 ± 0.00	**0.819 ± 0.02**	0.541 ± 0.03	0.651 ± 0.02
GB	0.200 ± 0.00	0.769 ± 0.01	0.532 ± 0.03	0.628 ± 0.02
MLP	0.192 ± 0.00	0.740 ± 0.02	0.681 ± 0.02	0.709 ± 0.00

e) TFIDF-1,2-STOPWORDS-LEMMA. Features: 129,745

Model	FPR	PRE	REC	F1
LR	0.193 ± 0.01	0.743 ± 0.02	0.662 ± 0.03	0.699 ± 0.02
NB	0.224 ± 0.00	**0.879 ± 0.03**	0.238 ± 0.01	0.374 ± 0.01
LSVM	0.197 ± 0.01	0.716 ± 0.02	0.749 ± 0.02	**0.732 ± 0.01**
SGD	0.204 ± 0.00	0.704 ± 0.01	**0.758 ± 0.01**	0.730 ± 0.01
SVM	**0.187 ± 0.01**	0.764 ± 0.02	0.637 ± 0.03	0.694 ± 0.02
KNN	0.253 ± 0.01	0.642 ± 0.01	0.656 ± 0.02	0.649 ± 0.01
RF	**0.187 ± 0.01**	0.841 ± 0.02	0.482 ± 0.05	0.611 ± 0.04
GB	0.200 ± 0.00	0.767 ± 0.01	0.538 ± 0.03	0.631 ± 0.02
MLP	0.203 ± 0.01	0.709 ± 0.01	0.733 ± 0.01	0.721 ± 0.01

f) TFIDF-1,2,3-STOPWORDS-LEMMA. Features: 261,665

Model	FPR	PRE	REC	F1
LR	0.202 ± 0.01	0.723 ± 0.02	0.675 ± 0.03	0.698 ± 0.02
NB	0.228 ± 0.01	0.672 ± 0.01	0.741 ± 0.02	0.704 ± 0.01
LSVM	0.211 ± 0.01	0.692 ± 0.02	0.778 ± 0.02	**0.732 ± 0.01**
SGD	0.219 ± 0.01	0.680 ± 0.02	**0.787 ± 0.01**	0.729 ± 0.01
SVM	**0.193 ± 0.01**	0.751 ± 0.02	0.639 ± 0.03	0.690 ± 0.02
KNN	0.253 ± 0.01	0.641 ± 0.01	0.656 ± 0.01	0.648 ± 0.01
RF	0.194 ± 0.01	**0.832 ± 0.01**	0.453 ± 0.05	0.585 ± 0.04
GB	0.201 ± 0.00	0.768 ± 0.01	0.531 ± 0.02	0.628 ± 0.01
MLP	0.203 ± 0.00	0.704 ± 0.01	0.760 ± 0.01	0.731 ± 0.00

Table 5: Best general results.

Placing	Experiment	Vocabulary	FPR	PRE	REC	F1
1	TFIDF-1,2-STOPWORDS-LEMMA-LSVM	129745	0.197	0.717	0.750	0.733
2	TFIDF-1,2,3-STOPWORDS-LEMMA-LSVM	261665	0.211	0.692	0.778	0.733
3	TFIDF-1,2,3-STOPWORDS-LEMMA-MLP	261665	0.204	0.705	0.761	0.731
4	TFIDF-1,2,3-SGD	384783	0.212	0.692	0.777	0.731
5	TFIDF-1,2-STOPWORDS-LEMMA-SGD	129745	0.204	0.704	0.759	0.730
6	TFIDF-1,2,3-STOPWORDS-LEMMA-SGD	261665	0.219	0.681	0.787	0.730
7	TFIDF-1,2,3-LSVM	384783	0.210	0.696	0.768	0.729
8	BOW-1,2-STOPWORDS-LEMMA-MLP	129745	0.173	0.781	0.679	0.726
9	TFIDF-1,2-MLP	156182	0.197	0.722	0.731	0.726
10	TFIDF-1,2-LSVM	156182	0.198	0.721	0.729	0.724

differences, since the Fake.Br corpus is composed of only text in journalistic style collected from websites, while in the FakeWhatsApp.Br the texts are predominantly short and stylistic varied, containing not only news, but also rumors, satirical and humorous texts, political propaganda, and others. In the following Subsection we analyze the failures of one of the best models.

5.1 Error Analysis

As described in Subsection 4.2, to take a deeper look at the limitations of our approach, we retrained and evaluated one of the best combination of features and method using a randomly separated train and test sets in a 80%-20% proportion. We used the LSVM method with TF-IDF vectors, unigrams, bigrams and trigrams, stopwords removal and lemmatization.

The test set has 1057 instances, of which 618 (58.4%) are negative (non-misiformation) and 439 (41.6%) are positive (misinformation). The resulting confusion matrix of the classification is presented in Table 6. We conducted a qualitative analysis of the 246 errors, formed by 105 (43% of total errors) false negative, that is, misinformation erroneously classified as non-misinformation, and by 141 (57% of total errors) false positives, non-misinformation erroneously classified as misinformation. We consider the false negatives more critical in this context, since the goal of a automated detection system would be alert human users, that could do their own fact-checking and reach a conclusion. So, a false positive could be later proven as such, but a false negative may not be taken in consideration for fact-checking.

Table 6: Confusion matrix for the test with LSVM.

	Predicted class			
Actual class		0	1	Total
	0	True Negative: 477 (45.13%)	False Positive: 141 (13.34%)	618
	1	False Negative: 105 (9.93%)	True Positive: 334 (31.60%)	439
	Total	582	475	1057

We categorized the texts wrongly classified in the following types, with examples and it's translations for English:

- **Short Text with External Information:** a short text that is followed by a media file (image, audio or video), or a URL to a Web page. As most of useful information in the message is not in the text itself, it's difficult for a pure NLP approach to detect a pattern to make a correct classification;

  False Negative E.g.: "Escuta a fala sensata e inteligente do Miguel Falabella"[10]

  Translation: "Listen to Miguel Falabella's sensible and intelligent speech."

- **Short Text:** in the case of a false negative, it is a short claim with a false allegation without a additional source of information. However, as happens in the previous type, the classifier may be biased to the size of text. In case of the false positives, it's short text with a opinion or a alert that it's not untrue but the use of alarmist words may be misleading;

  False Negative E.g.: "As Operadoras *Tim* a *Claro* e a *Oi* fazem 26 anos hoje. Envie isto para 20 pessoas, em seguida olhe seu saldo no *222/544/805* e você ganha *R$900,00* em créditos válidos por *120* meses. Funciona mesmo acabou de cair no meu"[11]

  Translation: "Cell Phone Operators Tim, Claro and Oi turn 26 today. Send this to 20 people, then look at your balance at 222/544/805 and you earn R$ 900.00 in credits valid for 120 months. Actually works, just happens with me."

- **Others:** This broad category includes long texts of different types. The false negatives can be opinions, satires or rumors, which mix true information with incorrect, inaccurate or extremely biased information. Although this class contains more textual information, the similarity with opinions labeled as non-misinformation may lead the model to the error. For the false positive, may be opinion or humorous texts, prayers or political propaganda, with a linguistic style that resembles misinformation messages;

  False Negative E.g.: "Gente Apenas Minha Opinião então Vamos lá. No Dia 06 de Junho TSE Derruba o Voto Impresso de Autoria do deputado Federal Candidato a presidente Jair Bolsonaro. No Dia 06 de Setembro Jair Bolsonaro Sofre um Atentado que Seria pra MATAR. Um Dia Antes da Eleição Dia 06; Coincidência Se Juntar as Datas dar Certos 666. Agora Bolsonaro corre novo Risco... A que Interessa Isso ? Nova Ordem Mundial? Marconaria ? iluminati ? Peço que Compartilhem E Faca Chegar ao Bolsonaro Breno Washington MG Juntos somos fortes"

  Translation: "Guys, this is just my opinion so let's go. On June 6, the paper vote proposal by the federal deputy candidate for president Jair Bolsonaro is overturned. On September 6, Jair Bolsonaro suffers an attack that was supposed to kill him. One day before the election on the 6th; coincidence joining dates gives 666. Now Bolsonaro is at new risk... to whom it matters? New world order? Masonry? Iluminati? I ask you to share and make it to the Bolsonaro. Breno Washington MG. Together we are strong"

Table 7: Percentage of each kind of error in false negatives and in false positives.

Type	**Percentage of false negative**	**Percentage of false positive**
Short text with external information	**71.2%**	**59.6%**
Short text	20.2%	17%
Other	8.6%	23.4%

The proportion of each type of text is shown in Table 7. We see from this Table that short texts with external information were the main cause of errors of both types, but it was more critical for false negatives. The short texts were the second type with more false negative, while the others type was a minority of false negative. However, the contrary happens when we look to the false positive, where the other type of texts are in second place and the short texts are a minority.

This results indicate that the model is in fact biased to the size of the text, tending to classify long texts as misinformation and short texts as non-

[10] https://g1.globo.com/fato-ou-fake/noticia/2018/10/26/e-fake-que-miguel-falabella-gravou-audio-sobre-cenario-pos-eleicao.ghtml

[11] https://www.boatos.org/tecnologia/tim-claro-oi-creditos-gratis.html

misinformation. It's necessary to take in consideration that short texts with external information can be very challenging to classify with a BoW approach alone, since short texts are usually noisier, less topic-focused and do not provide enough word co-occurrence or shared context. Therefore, machine learning methods that rely on the word frequency usually fail to achieve desire accuracy due to the data sparseness (Song et al., 2014).

As a last experiment, we select only the messages with 50 or more words, a subset of 1555 messages (29.4% of the original dataset), and repeat the same train and test procedure used in this subsection. As expected, performance increases significantly, achieving a F1 score of 0.87, a recall of 0.95, a precision of 0.80, and a false positive rate of 0.22. However, this is a change of the original task, since short messages are majority in the context of WhatsApp messages. Specific strategies must be developed for that issue.

6 CONCLUSIONS

The fast spread of misinformation through WhatsApp messages poses as major social problem. In this work, we presented a large-scale, labelled and public dataset of WhatsApp messages in Brazilian Portuguese. In addition, we performed a wide set of experiments seeking out to build a solution to the misinformation detection problem, in this specific context. Our findings help us to answer the research questions:

1. *How challenging is the task of misinformation detection in WhatsApp messages using NLP-based techniques?* We experimented a varied combination of BoW features and machine learning classification methods, resulting in a total of 108 combinations, and performed 5-fold cross validation in each combination. Our best results achieved a F1-score of 0.733, which may serve as a baseline for future work. The results shows that trustful misinformation detection in WhatsApp messages is still a open problem.

2. *Which combination of pre-processing methods, word-level features and classification algorithms are best suited for this task?* Our experiments showed that the methods MLP, LSVM and SGD achieved the best results in nearly every scenario of features. For TF-IDF vectors, the results improved when were used unigrams, bigrams and trigrams as tokens, stopwords removal and lemmatization, which was the best scenario.

3. *Which are the limitations of an NLP-based approach?* Finally, the qualitative analysis of the errors of our best results showed that the majority of errors occurred in short texts that refers to a media file or a website, thus resulting in a lack of information for the model and limiting the performance of a pure NLP-based approach. This analysis also indicates that the model was biased to classify long messages as misinformation, due the average difference in size of the two classes. When we filtered short texts from the dataset and repeated the classification experiment, the performance improved substantially, with a F1-score of 0.87.

In future work, we intend to investigate how the metadata associated with the message (senders, timestamps, groups where it was shared, etc) can be combined with textual features to improve classification. We also intend to investigate the task of multi-modal misinformation detection, extracting features from text and media files using a Deep Leaning approach. Finally, as misinformation varies over time, we intend to investigate semi-automatic methods for building continuously labeled WhatsApp's datasets.

ACKNOWLEDGEMENTS

This study was financed in part by the CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

REFERENCES

Faustini, P. and Covões, T. (2019). Fake news detection using one-class classification. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 592–597.

Gaglani, J., Gandhi, Y., Gogate, S., and Halbe, A. (2020). Unsupervised whatsapp fake news detection using semantic search. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 285–289. IEEE.

Garimella, K. and Tyson, G. (2018). Whatsapp, doc? a first look at whatsapp public group data. *arXiv preprint arXiv:1804.01473*.

Granik, M. and Mesyura, V. (2017). Fake news detection using naive bayes classifier. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 900–903. IEEE.

Guo, B., Ding, Y., Yao, L., Liang, Y., and Yu, Z. (2019). The future of misinformation detection: New perspectives and trends.

Kim, S.-B., Han, K.-S., Rim, H.-C., and Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11):1457–1466.

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

Maalej, Z. (2001). *Discourse Studies*, 3(3):376–378.

Machado, C., Kira, B., Narayanan, V., Kollanyi, B., and Howard, P. (2019). A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. WWW '19, page 1013–1019, New York, NY, USA. Association for Computing Machinery.

Monteiro, R. A., Santos, R. L., Pardo, T. A., De Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

Newman, N., Fletcher, R., Schulz, A., Andi, S., and Nielsen, R.-K. (2020). Reuters institute digital news report 2020. *Report of the Reuters Institute for the Study of Journalism*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Soetrisno, Y. A. A., Arfan, M., and Sofwan, A. (2017). Hoax detection system on indonesian news sites based on text classification using svm and sgd. In *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 45–49. IEEE.

Qiu, X., Oliveira, D. F., Shirazi, A. S., Flammini, A., and Menczer, F. (2017). Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):0132.

Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures.

Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, page 387–390, New York, NY, USA. Association for Computing Machinery.

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., and Kraus, S. (2018). A study of whatsapp usage patterns and prediction models without message content. *arXiv preprint arXiv:1802.03393*.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. (2014). Short text classification: A survey. *Journal of multimedia*, 9(5):635.

Su, Q., Wan, M., Liu, X., and Huang, C.-R. (2020). Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research*, 1:1–13.

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

Waterloo, S. F., Baumgartner, S. E., Peter, J., and Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. *new media & society*, 20(5):1813–1831.

Zervopoulos, A., Alvanou, A. G., Bezas, K., Papamichail, A., Maragoudakis, M., and Kermanidis, K. (2020). Hong kong protests: Using natural language processing for fake news detection on twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 408–419. Springer.