Detection of Misinformation about COVID-19 in Brazilian Portuguese WhatsApp Messages*

Antônio Diogo Forte Martins[1], Lucas Cabral[1], Pedro Jorge Chaves Mourão[2], José Maria Monteiro[1], and Javam Machado[1]

[1] Federal University of Ceará, Fortaleza, Ceará, Brazil
{diogo.martins, jose.monteiro, javam.machado}@lsbd.ufc.br
lucascabral@aridalab.dc.ufc.br
[2] Universidade Estadual do Ceará, Fortaleza, Ceará, Brazil
pedro.mourao@aluno.uece.br

Abstract. During the coronavirus pandemic, the problem of misinformation arose once again, quite intensely, through social networks. In many developing countries such as Brazil, one of the primary sources of misinformation is the messaging application WhatsApp. However, due to WhatsApp's private messaging nature, there still few methods of misinformation detection developed specifically for this platform. Additionally, a MID model built to Twitter or Facebook may have a poor performance when used to classify WhatsApp messages. In this context, the automatic misinformation detection (MID) about COVID-19 in Brazilian Portuguese WhatsApp messages becomes a crucial challenge. In this work, we present the COVID-19.BR, a data set of WhatsApp messages about coronavirus in Brazilian Portuguese, collected from Brazilian public groups and manually labeled. Besides, we evaluated a series of misinformation classifiers combining different techniques. Our best result achieved an F1 score of 0.778, and the analysis of errors indicates that they occur mainly due to the predominance of short texts. When texts with less than 50 words are filtered, the F1 score rises to 0.857.

Keywords: Misinformation Detection · Fake News Detection · Natural Language Processing · WhatsApp · COVID-19

1 Introduction

During the coronavirus pandemic, the problem of misinformation arose once again, quite intensely, through social networks. In April 2020, the United Nations (UN) declared that there is a "dangerous misinformation epidemic", responsible for the spread of harmful health advice and false solutions. The misinformation concept can be understood as a process of intentional production of a communicational environment based on false, misleading, or decontextualized information to cause a communicational disorder [11].

* Supported by CAPES and LSBD.

Currently, the main tool used to spread misinformation is WhatsApp instant messaging application. Through this application, misinformation can deceive thousands of people in a short time, bringing great harm to public health. A very relevant WhatsApp feature is the public groups which are accessible through invitation links published on popular websites and social networks. Each group can put together a maximum of 256 members and they usually have specific topics for discussion, very similar to social networks. Thus, these public groups have been used to spread misinformation.

In this context, the automatic misinformation detection (MID) about COVID-19 in Brazilian Portuguese WhatsApp messages becomes a crucial challenge. In a wide definition, MID is the task of assessing the appropriateness (truthfulness, credibility, veracity, or authenticity) of claims in a piece of information [11]. However, due to WhatsApp's private messaging nature, there are still few MID methods developed specifically for this platform. Additionally, a MID model built to Twitter or Facebook may have a poor performance when used to classify WhatsApp messages. A model's performance is highly dependent on the linguistic patterns, topics, and vocabulary present in the data used to train it. Nevertheless, the linguistic patterns found in WhatsApp messages are quite different from those found in Facebook and Twitter [12]. Thus, despite the scientific community's efforts, there is still a need for a large-scale corpus containing WhatsApp messages in Portuguese about COVID-19.

In order to fill this gap, we built a large-scale, labeled, anonymized, and public data set formed by WhatsApp messages in Brazilian Portuguese (PT-BR) about coronavirus pandemic, collected from public WhatsApp groups. Then, we conduct a series of classification experiments using different machine learning methods to build an efficient MID for WhatsApp messages. Our best result achieved an F1 score of 0.778 due to the predominance of short texts.

2 Related Work

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

The study presented in [2] proposes a misleading-information detection model that relies on several contents about COVID-19 collected from the World Health Organization, UNICEF, and the United Nations, as well as epidemiological material obtained from a range of fact-checking websites. The research presented in [1] proposed a set of machine learning techniques to classify information and misinformation. In [6], the authors introduced CoVerifi, a web application that combines both the power of machine learning and the power of human feedback to assess the credibility of news about COVID-19. The study presented in [4] proposed a multimodal multi-image system that combines information from different modalities in order to detect fake news posted online.

3 Data Set Design

An important aspect to consider while developing a MID method for WhatsApp messages in Brazilian Portuguese is the necessity of a large-scale labeled data set. However, there is no corpus for Brazilian Portuguese with these characteristics as far as we know. Besides, due to its private chat purpose, WhatsApp does not provide a public API to automatically collect data. Thus, build this data set is a technical, also ethical challenge. For this reason, we used a methodology similar to [10, 3] to build a large-scale labelled corpus of WhatsApp messages in Brazilian Portuguese.

In order to create the data set presented in this paper, we collected messages from open WhatsApp groups. These groups were found by searching for "chat.whatsapp.com/" on the Web. Next, we analyzed the theme and purpose of each group found previously. Then, we selected 236 public groups. After this, we joined these groups and started collecting messages. Each collected message is stored in a row of the data set. Finally, we select a message subset called "viral messages". We defined "viral messages" as identical messages with more than five words that appear more than once in the data set. It is important to highlight that sensitive attributes such as user name, cell phone number and group name were anonymized using hash functions. Figure 1 shows an extract from our data set after anonymization and before data labeling. Our data set has 228061 WhatsApp messages from users and groups from all over Brazil.

	id	date	hour	ddi	country	country_iso3	ddd	state	group	midia	url	characters	words	viral	sharings	text
0	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	9	1	0	1	Morreram?
1	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	24	4	0	1	Olá novato, se apresente
2	5788106393468158140	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	9	2	0	1	há tempos
3	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	13	2	0	1	Leglão Urbana
4	5788106393468158140	07/04/20	04:13	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	6	1	0	1	Indios

Fig. 1. Extract from the collected data before the labeling process.

In order to build a high-quality corpus, data labelling is another hard challenge since we have to specify if the text is true or false based on trusted sources, such as specialized journalists or fact-checking sites. So, we conducted the data labeling process entirely manually. A human specialist checked each message's content and determined if it contains or not misinformation. Since this process is time-consuming, we chose to label only unique messages containing the following keywords: *"covid", "coron", "virus", "china", "chines", "cloroquin", "vacina"*. The resulting data set now has 2899 unique messages. We labeled all these messages with the general misinformation definition adopted in [11] labeling them as 0 if the message does not contain misinformation and 1 if it contains misinformation. Three annotators, two computer science masters students and one sociologist, conducted the labeling process. We solved labeling disagreements executing a collective review round.

Our labeling process was based on the following steps. If the text contains verifiable untrue claims, we annotate it as misinformation. We made use of trustful Brazilian fact-checking platforms such as *Agência Lupa*[3] and *Boatos.org*[4]. If the text contains imprecise, biased, alarmist, or harmful claims that cannot be proven, we annotate it as misinformation. If the text is short and accompanied by media content (image, video, or audio), we search on the web for the media content and, if we find the corresponding media, we decide the label based on the previous criteria. If the original media cannot be found, we use the second criterion to label it. And If none of the previous criteria is found in the text, we label it as not containing misinformation.

After the labeling process, we removed messages with only *url* as text content. So, the resulting corpus contains 532 unique messages labeled as misinformation (label 1) and 858 unique messages labeled as non-misinformation (label 0). Table 1 presents basic statistics about the data set.

Table 1. Data set basic statistics.

Statistics	Non-misinformation	Misinformation
Count of unique messages	858	532
Mean and std. dev. of number of tokens	92.02 ± 203.24	167.02 ± 248.02
Minimum number of tokens	1	1
Median number of tokens	20	50
Maximum number of tokens	3100	1666
Mean and std. dev. of shares	2.51 ± 4.85	2.47 ± 3.41

4 Experiments

We have explored multiple combinations between feature extraction from text and classification algorithms. We performed our experiments using k-fold cross-validation with $k = 5$ folds. We also performed a Bayesian optimization over hyperparameters to search the optimal configuration for the best classifiers. Besides, we evaluate different techniques for text feature extraction, but we decided to use traditional Bag-Of-Words (BoW) and TF-IDF text representations in our experiments. Since one of our goals is to define a baseline for automatic MID about the COVID19 in WhatsApp messages in Brazilian Portuguese, these techniques features are suitable for this purpose and have been already used in a wide range of text classification problems.

Our text pre-processing method consists in convert to lowercase, separate emojis with white spaces to avoid generating a new token for each emoji sequence, and maintain only the domain name for *urls*. Because of the lexical diversity of the corpus, the resulting vectors have large dimensions and sparsity. Moreover, we added more variety to our experiments by using different n-gram values. So, we combined these different vectorization techniques (TF-IDF or binary BoW), the n-grams range (unigrams, bigrams, and trigrams), and the

[3] http://piaui.folha.uol.com.br/lupa/
[4] http://www.boatos.org/

extra steps of pre-processing (lemmatization and stop words removal), leading to a total of 12 different feature extraction scenarios.

For each scenario, we performed experiments using nine machine learning classification techniques, already used in several text classification tasks [7]: logistic regression (LR), Bernoulli (if the features are BoW) or Complement Naive-Bayes (if features are TF-IDF) (NB) [5, 9], support vector machines with a linear kernel (LSVM), SVM trained with stochastic gradient descent (SGD), SVM trained with an RBF kernel [8] (SVM), K-nearest neighbors (KNN), random forest (RF), gradient boosting (GB), and multilayer perceptron neural network (MLP). At first, all techniques were used with default hyperparameters. Next, we performed a Bayesian optimization to find the optimal hyperparameters for the best combinations of features and classifiers.

Just considering all combinations between features, pre-processing, and classification methods and excluding the Bayesian optimization step, we performed a total of 108 experiments, all of them using k-fold cross-validation with $k = 5$.

In order to evaluate the performance of the experiments and considering we are working with a binary classification task, where non-misinformation represents the negative class and misinformation the positive, we use the following metrics: False positive rate (FPR), Precision (PRE), Recall (REC), and F1-score (F1). Because we use k-fold cross-validation, each metric's mean are collected and will also be presented.

5 Results

For the sake of readability, we included only the results of the top 10 best combinations of classifiers and features extraction techniques. The results presented in the following tables are the metrics' mean after 5 rounds of k-fold cross-validation.

Table 2 summarizes the results for the experiments we run with standard hyperparameters. Analyzing the F1 values, we can observe that the difference is not large, less than 1% from the first to last. We achieved the best results when using BoW and NB. The removal of stop words and lemmatization helped improve some of NB results in the trigram and bigram scenarios. When using TF-IDF and LSVM, we achieved the lowest value of FPR among the top 5 results. The best result was obtained using BoW as feature extractor, bigram, removing stop words and performing lemmatization, and with the NB classifier.

Next, we performed a Bayesian optimization over the hyperparameters to search the optimal configuration for the classifiers. For NB, the best value of *alpha* was 0; for LSVM, the best value of C was 348.61; for SGD, the best value of *alpha* was 0.00185. Table 3 summarizes the results of the experiments with the best hyperparameters. Analyzing the results, we can see that now the best combination of classifier and features extraction techniques is SGD using BoW as feature extractor, trigram, removing stop words, and performing lemmatization (with 0.4% of improvement in F1 and 3.3% of improvement in FPR). Besides, the result shows that, even if we searched for hyperparameters for a specific

Table 2. Top 10 best combinations of classifiers and features extraction techniques. All presented metrics values are the mean after 5 rounds of cross-validation.

Rank	Experiment	Vocabulary	FPR	PRE	REC	F1
1	BOW-BIGRAM-LEMMA-NB	70986	0.179	0.734	0.840	0.774
2	TFIDF-BIGRAM-LSVM	84189	0.149	0.775	0.780	0.773
3	BOW-UNIGRAM-NB	15165	0.183	0.734	0.833	0.771
4	TFIDF-TRIGRAM-SGD	190376	0.160	0.746	0.804	0.770
5	BOW-TRIGRAM-LEMMA-NB	147900	0.182	0.728	0.836	0.770
6	BOW-UNIGRAM-LEMMA-NB	13039	0.183	0.730	0.836	0.769
7	TFIDF-TRIGRAM-LEMMA-SGD	147900	0.162	0.741	0.808	0.769
8	BOW-BIGRAM-NB	84189	0.181	0.733	0.827	0.768
9	BOW-TRIGRAM-NB	190376	0.178	0.736	0.821	0.768
10	TFIDF-TRIGRAM-MLP	190376	0.152	0.779	0.772	0.768

combination, we improved the SGD classifier performance using different feature extraction methods.

Table 3. Top 10 best combinations of classifiers and features extraction using the Bayesian optimization hyperparameters. All presented metrics values are the mean after 5 rounds of cross-validation.

Rank	Experiment	Vocabulary	FPR	PRE	REC	F1
1	BOW-TRIGRAM-LEMMA-SGD	147900	0.146	0.771	0.791	0.778
2	BOW-BIGRAM-LEMMA-NB	70986	0.179	0.734	0.840	0.774
3	BOW-UNIGRAM-NB	15165	0.183	0.734	0.833	0.771
4	BOW-TRIGRAM-LEMMA-NB	147900	0.182	0.728	0.836	0.770
5	BOW-UNIGRAM-LEMMA-NB	13039	0.183	0.730	0.836	0.769
6	BOW-BIGRAM-NB	84189	0.181	0.733	0.827	0.768
7	BOW-TRIGRAM-NB	190376	0.178	0.736	0.821	0.768
8	TFIDF-BIGRAM-LEMMA-LSVM	70986	0.159	0.755	0.789	0.766
9	TFIDF-BIGRAM-LEMMA-MLP	70986	0.158	0.765	0.772	0.763
10	TFIDF-BIGRAM-MLP	84189	0.157	0.778	0.756	0.760

Lastly, we decided to select only the messages containing 50 or more words from our data set, resulting in a subset of 269 messages with misinformation and 292 messages without misinformation. We repeated all the experiments to analyze the influence of the text length in the prediction. Table 4 shows the results for these experiments. We had a significant performance increase in this scenario, achieving an F1 of 0.857 when using BoW, unigram, and NB as the combination of features and classifier. In terms of FPR, we achieved a result of 0.14 using BoW, bigram, and MLP. By analyzing these results, we can observe that the text length affects the classifiers' performance since there are short messages in our data set linked to external media that contain misinformation.

From our results, we can recognize how difficult it is to perform MID in WhatsApp since our best result was an F1 of 0.778. When considering only long texts, our best F1 result is 0.857.

Table 4. Top 10 best combinations of classifiers and features extraction for long texts. All presented metrics values are the mean after 5 rounds of cross-validation.

Rank	Experiment	Vocabulary	FPR	PRE	REC	F1
1	BOW-UNIGRAM-NB	14186	0.153	0.846	0.885	0.857
2	BOW-BIGRAM-MLP	77174	0.140	0.862	0.862	0.856
3	BOW-BIGRAM-NB	77174	0.163	0.833	0.892	0.855
4	BOW-TRIGRAM-NB	173315	0.163	0.836	0.888	0.854
5	TFIDF-TRIGRAM-MLP	173315	0.156	0.831	0.888	0.853
6	BOW-BIGRAM-LEMMA-NB	64803	0.168	0.826	0.896	0.852
7	BOW-TRIGRAM-LEMMA-NB	134067	0.172	0.822	0.892	0.848
8	TFIDF-BIGRAM-LSVM	77174	0.169	0.820	0.881	0.844
9	TFIDF-UNIGRAM-LEMMA-MLP	12255	0.176	0.790	0.907	0.842
10	BOW-UNIGRAM-LR	14186	0.170	0.832	0.866	0.841

6 Conclusions

In this work, we presented a large-scale, labeled, and public data set of WhatsApp messages in Brazilian Portuguese about coronavirus pandemic. In addition, we performed a wide set of experiments seeking out to build an efficient solution to the MID problem in this specific context. Our best result achieved an F1 score of 0.778 due to the predominance of short texts. However, when texts with less than 50 words are filtered, the F1 score rises to 0.857. In future work, we pretend 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 our MID solution's performance. All the experiments and the COVID-19.BR data set are available at our public repository[5].

References

1. Choudrie, J., Banerjee, S., Kotecha, K., Walambe, R., Karende, H., Ameta, J.: Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. Computers in Human Behavior **119**, 106716 (2021). https://doi.org/10.1016/j.chb.2021.106716
2. Elhadad, M.K., Li, K.F., Gebali, F.: Detecting misleading information on covid-19. IEEE Access **8**, 165201–165215 (2020). https://doi.org/10.1109/ACCESS.2020.3022867
3. Garimella, K., Tyson, G.: Whatsapp, doc? a first look at whatsapp public group data. arXiv preprint arXiv:1804.01473 (2018)
4. Giachanou, A., Zhang, G., Rosso, P.: Multimodal multi-image fake news detection. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 647–654 (2020). https://doi.org/10.1109/DSAA49011.2020.00091
5. Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering **18**(11), 1457–1466 (2006)

[5] https://gitlab.com/jmmonteiro/misinformation_covid19

6. Kolluri, N.L., Murthy, D.: Coverifi: A covid-19 news verification system. Online Social Networks and Media **22**, 100123 (2021). https://doi.org/10.1016/j.osnem.2021.100123
7. Pranckevičius, T., Marcinkevičius, V.: 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 (2017)
8. Prasetijo, A.B., Isnanto, R.R., Eridani, D., Soetrisno, Y.A.A., Arfan, M., Sofwan, A.: 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). pp. 45–49. IEEE (2017)
9. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 616–623 (2003)
10. Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., Benevenuto, F.: A system for monitoring public political groups in whatsapp. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web. p. 387–390. WebMedia '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3243082.3264662
11. Su, Q., Wan, M., Liu, X., Huang, C.R.: Motivations, methods and metrics of misinformation detection: An nlp perspective. Natural Language Processing Research **1**, 1–13 (2020). https://doi.org/10.2991/nlpr.d.200522.001
12. Waterloo, S.F., Baumgartner, S.E., Peter, J., Valkenburg, P.M.: Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. new media & society **20**(5), 1813–1831 (2018)