Automatic Misinformation Detection About COVID-19 in Brazilian Portuguese WhatsApp Messages

Antônio Diogo Forte Martins[1], **José Maria Monteiro**[1], **Javam Machado**[1]

[1]LSBD – Computer Science Department – Federal University of Ceará

{diogo.martins, jose.monteiro, javam.machado}@lsbd.ufc.br

Nível: Mestrado
Ingresso: Março de 2020
Previsão de Término: Janeiro de 2023
Etapas já concluídas: Revisão Bibliográfica Preliminar, Definição do Problema
Defesa da Pré-Proposta: Agosto de 2021
Defesa da Proposta: Dezembro de 2021
Lista de Publicações:

- Detection of Misinformation About COVID-19 in Brazilian Portuguese WhatsApp Messages - NLDB2021
- LGPD: A Formal Concept Analysis and its Evaluation - SBBD 2020

Abstract. *During the coronavirus pandemic, the problem of misinformation arose once again, quite intensely, through social networks. In 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. 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. Then, we are investigating different machine learning methods in order to build an efficient MID for WhatsApp messages. So far, our best result achieved an F1 score of 0.774 due to the predominance of short texts. However, when texts with less than 50 words are filtered, the F1 score rises to 0.85.*

Resumo. *Durante a pandemia do coronavírus, o problema da desinformação voltou a surgir, de forma bastante intensa, nas redes sociais. No Brasil, uma das principais fontes de desinformação é o aplicativo de mensagens WhatsApp. No entanto, devido à natureza de mensagens privadas do WhatsApp, ainda existem poucos métodos de detecção de desinformação desenvolvidos especificamente para esta plataforma. Nesse contexto, a detecção automática de desinformação (MID) sobre o COVID-19 em mensagens do WhatsApp em português do Brasil torna-se um desafio crucial. Neste trabalho, apresentamos o COVID-19.BR, um conjunto de dados de mensagens do WhatsApp sobre coronavírus em português do Brasil, coletados de grupos públicos brasileiros e rotulados manualmente. Então, estamos investigando diferentes métodos de aprendizado de máquina para construir um MID eficiente para mensagens do WhatsApp. Até o momento, nosso melhor resultado foi de 0,774 na F1 devido ao predomínio de textos curtos. No entanto, quando textos com menos de 50 palavras são filtrados, a pontuação F1 sobe para 0,85.*

1. Introduction

Misinformation is a problem in our connected society and during the coronavirus pandemic it arose intensely through social networks one more time. In February 2020, the Brazilian Health Ministry reported that among 6,500 messages received and analyzed by it, between January 22 and February 27, 90% were related to the new virus. From the messages about coronavirus, 85% were false [1]. Misinformation is a process of intentional production of a communicational environment based on false, misleading, or decontextualized information in order to create a communicational disorder [Su et al. 2020].

WhatsApp is the favorite tool of misinformation spreaders. Through it, messages containing misinformation can reach thousands of people in a short period of time, bringing great harm to public health. WhatsApp's public groups which are easily accessible through invitation links available on popular websites and social networks is a relevant feature of the instant messaging application. These groups can have up to 256 members and, usually, they have an specific discussion topic, similar to social networks. Misinformation is massively spread in these public groups.

In this context, the automatic misinformation detection (MID) about COVID-19 in Brazilian Portuguese WhatsApp messages becomes a crucial challenge. MID is the task of assessing the appropriateness (truthfulness, credibility, veracity, or authenticity) of claims in a piece of information [Su et al. 2020]. In spite of that, due to WhatsApp's private messaging nature, there are not a wide variety of MID methods developed for this application. Also, a MID model built for Twitter of Facebook may not perform well classifying WhatsApp Messages, because the model's performance is highly dependent on the linguistic patterns presented in the data used during the training phase. WhatsApp messages are different from those found in Facebook and Twitter [Waterloo et al. 2018]. Each language has its specificity as well which means that, even if there is an efficient MID for WhatsApp in a language different than PT-BR, it will not classify correctly the messages.

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 are investigating different machine learning methods in order to build an efficient MID for WhatsApp messages. So far, our best result achieved an F1 score of 0.774 due to the predominance of short texts. However, when texts with less than 50 words are filtered, the F1 score rises to 0.85.

2. Related Work

The study presented in [Elhadad et al. 2020] 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 [Choudrie et al. 2021] proposed a set of machine learning techniques to classify information and misinformation. In [Kolluri and Murthy 2021], 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

[1] Available in: https://www.saude.gov.br/fakenews. Accessed in: 25 April, 2020

in [Giachanou et al. 2020] proposed a multimodal multi-image system that combines information from different modalities in order to detect fake news posted online.

In our previous work [Martins et al. 2021], we introduced COVID19.BR, 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 to fill the gap of data sets focused on WhatsApp messages. We conduct a series of classification experiments using different machine learning methods to build an efficient MID for WhatsApp messages.

3. Building an Efficient MID for WhatsApp Messages

3.1. 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 was 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 [Resende et al. 2018, Garimella and Tyson 2018] to build the COVID19.BR data set, a large-scale, labeled, anonymized, and public corpus of WhatsApp messages in Brazilian Portuguese about coronavirus pandemic [Martins et al. 2021]. Table 1 presents basic statistics about the COVID19.BR data set. This corpus contains 532 unique messages labeled as misinformation (label 1) and 858 unique messages labeled as non-misinformation (label 0).

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

3.2. Experiments

We have explored multiple combinations between feature extraction from text and classification algorithms. More precisely, we used Bag-Of-Words (BoW) and TF-IDF as feature extraction methods. 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 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: [Pranckevičius and Marcinkevičius 2017]: logistic regression (LR), Bernoulli (if the features are BoW) or Complement Naive-Bayes (if features are TF-IDF) (NB)

Table 2. Best combinations of classifiers and features extraction techniques.

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

[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 an RBF kernel [Prasetijo et al. 2017] (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.

3.3. 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 Table 2 are the metrics' mean after 5 rounds of k-fold cross-validation. The best result was obtained using BoW as feature extractor, bigram, removing stop words and performing lemmatization, and with the NB classifier (F1 of 0.778). Next, in order to analyze the influence of the text length in the prediction we decided to select only the messages containing 50 or more words from the COVID19.BR data set, resulting in a subset of 269 messages with misinformation and 292 messages without misinformation. Then, we repeated all the experiments using this subset. Table 3 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.

3.4. Improvements

From the classification perspective, we want to explore deep learning architectures in order to improve MID performance. Recurrent Neural Networks (RNNs), in special Long-Short Term Memory Networks (LSTMs) [Hochreiter and Schmidhuber 1997], are the

Table 3. Best combinations of classifiers and features extraction for long texts.

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

best solution since they can capture dependency in a sequence, in our case a sequence of words, thus capturing the message's context. We want to use more advanced text embedding with transformers such as BERT [Devlin et al. 2018]. Figure 1 shows MIDeepBR, a deep learning architecture that we are developing for this task. We achieved F1 of 0.834 in the initial experiments. Message's metadata is also an approach we want to add as features to the MID in order to improve results.

We also want to explore the data set in a qualitative point of view. We want to characterize the messages by analyzing its content. We have temporal information in our data set about the messages. With this in mind, we want to explain how the messages spread through the groups, with this we can identify the misinformation spreaders and understand how they use WhatsApp to deceive people. We want as well to use eXplainable Artificial Intelligence (XAI) to understand how the classifiers are working so we

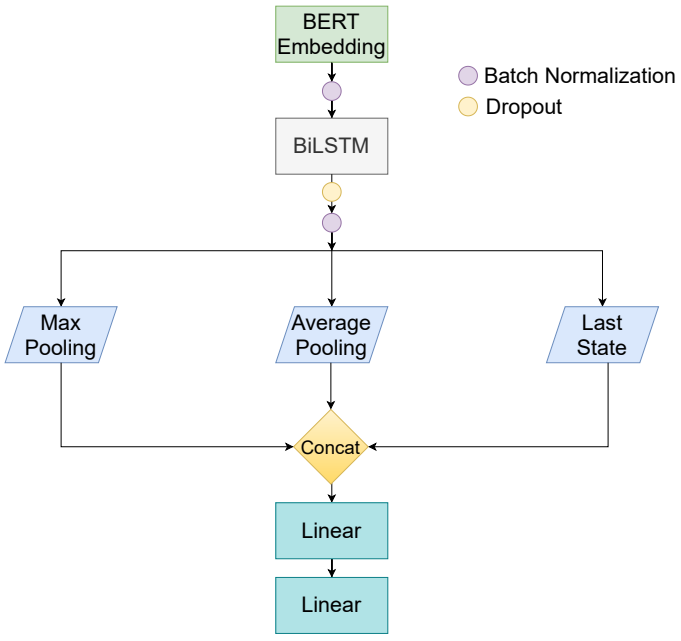

Figure 1. The MIDeepBR Architecture.

can analyze which are the most important words to state if a message contains or not misinformation.

4. Research Methodology

This research project consists in the following activities:

- State-of-the-art review of MID methods.
- Collect messages and metadata to build the data set.
- Label the misinformation data set.
- Develop a baseline with classic machine learning and NLP techniques.
- Develop a deep learning architecture using advanced NLP techniques as well to improve baseline results.
- Explain the models prediction using XAI tools.
- Characterize the messages by analyzing its content.
- Use message's metadata do explain how they spread through the groups.

5. Paper Submission Plan

We expect to submit and publish two more papers until July 2022 with the following improvements:

- Deep learning and advanced NLP techniques to improve the MID classification performance (SBBD 2021).
- Using message's metadata to explain how they spread through the WhatsApp groups and improve the MID classification performance (EDBT 2022).

6. Conclusion

In this work, we presented a large-scale, labeled, and public data set of WhatsApp messages in Brazilian Portuguese about coronavirus pandemic. Also, we present the obtained results from a wide set of experiments seeking out to build a baseline 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. To improve the MID performance, we want to use deep learning and advanced NLP techniques which we already obtained promising initial results of F1 score 0.834. From a qualitative point of view, we want to use the message's metadata do explain how they spread through the groups, characterize the messages, and explain the models predictions.

References

Choudrie, J., Banerjee, S., Kotecha, K., Walambe, R., Karende, H., and Ameta, J. (2021). Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in Human Behavior*, 119:106716.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elhadad, M. K., Li, K. F., and Gebali, F. (2020). Detecting misleading information on covid-19. *IEEE Access*, 8:165201–165215.

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

Giachanou, A., Zhang, G., and Rosso, P. (2020). Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Kolluri, N. L. and Murthy, D. (2021). Coverifi: A covid-19 news verification system. *Online Social Networks and Media*, 22:100123.

Martins, A. D. F., Cabral, L., Chaves Mourão, P. J., Monteiro, J. M., and Machado, J. (2021). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In Métais, E., Meziane, F., Horacek, H., and Kapetanios, E., editors, *Natural Language Processing and Information Systems*, pages 199–206, Cham. Springer International Publishing.

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.

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., 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.

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.

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.