Detection of Misinformation about COVID-19 in Brazilian Portuguese WhatsApp Messages Using Deep Learning

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

[1]Department of Computing, Federal University of Ceará, Fortaleza-Ceará, Brazil

[2]Universidade Estadual do Ceará, Fortaleza-Ceará, Brazil

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

lucascabral@aridalab.dc.ufc.br

pedro.mourao@aluno.uece.br

Abstract. *During the COVID-19 pandemic, the misinformation problem arose once again through social networks, like a harmful health advice and false solutions epidemic. In Brazil, as well as in many developing countries, one of the primary sources of misinformation is the messaging application WhatsApp. Thus, the automatic misinformation detection (MID) about COVID-19 in Brazilian Portuguese WhatsApp messages becomes a crucial challenge. Still, due to WhatsApp's private messaging nature, there are still few methods of misinformation detection developed specifically for the WhatsApp platform. In this paper, we propose a new approach, called MIDeepBR, based on BiLSTM neural networks, pooling operations and attention mechanism, which is able to automatically detect misinformation in Brazilian Portuguese WhatsApp messages. Experimental results evidence the suitability of the proposed approach to automatic misinformation detection. Our best results achieved an F1 score of 0.834, while in previous works, the best results achieved an F1 score of 0.778. Thus, MIDeepBR outperforms the previous works.*

1. Introduction

Misinformation is a major issue in our society, and unfortunately, during the coronavirus pandemics, it arose intensely through social networks. The United Nations (UN) stated in April 2020 that there is a "dangerous misinformation epidemic" responsible for disseminating misleading advice and solutions about the coronavirus[1]. 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[2].

The misinformation concept can be defined as a process of intentional production of a communicational environment based on false, misleading, or decontextualized information to cause a communicational disorder [Su et al. 2020]. Nevertheless, the term fake news, despite specifically describe intentionally misleading information written as

[1]UN. "Hatred going viral in 'dangerous epidemic of misinformation' during COVID-19 pandemic". 14 April, 2020. Available in: https://news.un.org/en/story/2020/04/1061682. Accessed on: April 25, 2020.

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

journalistic news, has become very present in popular culture and is sometimes used as a misinformation synonym [Guo et al. 2019].

WhatsApp instant messaging application is currently the main misinformation spread channel. WhatsApp is very popular in Brazil, with more than 120 million users in a population of about 210 million people [Resende et al. 2019]. In February 2020, the Panorama Mobile Time/Opinion Box survey on mobile messaging in Brazil revealed that WhatsApp is installed on 99% of Brazilian smartphones. Among users of the application, 98% said they access it every day or almost every day [3]. Through it, messages containing misinformation can mislead thousands of people in a short period bringing great harm to public health. A survey by the Oswaldo Cruz Foundation (Fiocruz) showed that 73.7% of the false news about the new coronavirus circulated through WhatsApp. Another 10.5% were published on Instagram and 15.8% on Facebook[4]. A very relevant WhatsApp feature is the public groups. These public groups are accessible through invitation links published on popular websites and social networks, such as Facebook and Twitter. Usually, they have specific topics for discussion, such as health, politics, and sports. Each group can put together a maximum of 256 members. So, WhatsApp public groups are very similar to social networks. Thus, they have been used to spread misinformation. Moreover, due to the high volume of information that we are exposed to, we have a limited ability to distinguish true information from misinformation [Vosoughi et al. 2018, Qiu et al. 2017].

In this context, the automatic misinformation detection (MID) about COVID-19 in Brazilian Portuguese WhatsApp messages becomes a crucial challenge. 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. 2020]. Early detection of misinformation could prevent its spread, thus reducing its damage. MID approaches have been extensively used with data collected from platforms like Facebook[5] [Granik and Mesyura 2017] and Twitter[6] [Zervopoulos et al. 2020]. However, MID models built using Twitter or Facebook data may perform poorly in classifying WhatsApp messages since the linguistic patterns of WhatsApp messages are different from those found in Facebook and Twitter[Waterloo et al. 2018, Rosenfeld et al. 2018]. The performance of a model for this kind of task is extremely dependent on the linguistic patterns and vocabulary present in the corpus used to train it. Nevertheless, due to the privacy requirements of WhatsApp, there are few methods specifically developed for it.

In our previous paper [Martins et al. 2021], we presented the COVID-19.BR, a large-scale, labelled, anonymized, and public data set formed by WhatsApp messages in Brazilian Portuguese (PT-BR) about coronavirus pandemic, collected from public WhatsApp groups using the platform proposed in [de Sá et al. 2021]. In that work, we conduct a series of classification experiments using nine different machine learning methods to build an efficient MID for WhatsApp messages: logistic regression (LR), Complement Naive-Bayes, support vector machines with a linear kernel (LSVM), SVM trained with

[3]SCHERMANN, Daniela. Panorama Mobile Time/Opinion Box: Mensageria no Brasil. Opinion Box, 2 mar. 2018. Available in https://blog.opinionbox.com/mensageria-no-brasil-sexta-edicao/. Accessed in: 11 mar. 2020.

[4]Available in: https://portal.fiocruz.br/noticia/pesquisa-revela-dados-sobre-fake-news-relacionadas-covid-19. Accessed in: 27 April, 2020.

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

[6]https://twitter.com/

stochastic gradient descent (SGD), SVM trained with an RBF kernel (SVM), K-nearest neighbors (KNN), random forest (RF), gradient boosting (GB), and multilayer perceptron neural network (MLP). The best result reached by [Martins et al. 2021] had an F1 score of 0.778, considering the full corpus of COVID-19.BR data set.

This paper proposes a new approach, called MIDeepBR, based on BiLSTM neural networks, pooling operations and attention mechanisms. MIDeepBR can automatically detect misinformation in PT-BR WhatsApp messages. MIDeepBR will automatically detect misinformation at the Digital Lighthouse [de Sá et al. 2021] platform. Experimental results evidence the suitability of the proposed approach to automatic misinformation detection. Our best results achieved an F1 score of 0.834, while in previous works [Martins et al. 2021], the best results achieved an F1 score of 0.778. Thus, MIDeepBR outperforms our previous work.

The remainder of this paper is organized as follows. Section 2 presents the main related work. Section 3 provides an overview of the theoretical background. Section 4 describes our deep learning approach, called MIDeepBR, to detect misinformation in WhatsApp messages about coronavirus in PT-BR. Section 5 details our experimental evaluation. Section 6 reports and discusses the results. Conclusions and future work are presented in Section 7.

2. Related Work

Divers 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, such as Twitter and Facebook, are amongst the main data sources used to build misinformation data sets.

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 authors use this collected ground-truth data to build a misinformation detection system. Ten machine learning algorithms, with seven feature extraction techniques, were used to construct a voting ensemble machine learning classifier. The research presented in [Choudrie et al. 2021] proposed a set of machine learning techniques to classify information and misinformation. They achieved a classification accuracy of 86.7% with the Decision Tree classifier and 86.67% with the Convolutional Neural Network model.

In [Kolluri and Murthy 2021], the authors introduced CoVerifi, a web application that combines the power of machine learning and human feedback to assess the credibility of news about COVID-19. By allowing users to "vote" on news content, the CoVerifi platform will allow the data labeling in an open and fast way.

In [Maakoul et al. 2020], the authors provide an aggregation system to detect and analyze fake news related to the COVID'19 pandemic in the Moroccan context based on data sets scrapped from Facebook. They approach the problem of fake news related to COVID'19 as a global pandemic. The study presented 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 particular, the system combines textual, visual, and semantic information.

Thus, despite the scientific community's efforts, there is still a need for new methods and approaches to automatic misinformation detection in PT-BR WhatsApp messages about COVID-19. 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. In this context, this paper's main contribution is a new approach, called MIDeepBR, based on BiLSTM neural networks, pooling operations and attention mechanism, which can automatically detect misinformation in PT-BR WhatsApp messages.

3. Theoretical Background

3.1. Long-short term memory

Recurrent Neural Networks (RNNs) are a particular type of neural networks that can efficiently handle sequential data. They are used when working with sequential dependencies. Their effectiveness in handling sequential inputs arises because they use the last neural network cell output as input to the next element of the sequence. Long-Short Term Memory (LSTMs) [Hochreiter and Schmidhuber 1997] is a special type of RNNs. It enhances the RNNs making use of complex gating mechanisms. These improvements allow the network to handle long sequences. A regular LSTM works by handle long sequences with the last elements. However, in some cases, the current element of the sequence may have a sequential dependency not only with past elements but also with future elements. Bidirectional LSTM [Graves and Schmidhuber 2005] addresses this problem by running the sequence into one LSTM performing the operations in the forward direction and other LSTM in the backward direction, and then concatenating the results.

3.2. Transformers

Transformer [Vaswani et al. 2017] is a very popular Natural Language Processing technique. It is a multi-layered deep learning architecture composed of encoding and decoding blocks. The encoder block has a self-attention [Vaswani et al. 2017] layer connected to a feed-forward neural network layer. The decoder block has the same layers, but it has another attention layer called encoder-decoder attention between the feed-forward and the self-attention. This encoder-decoder attention layer input is the last decoder block output and the output of the last encoder block.

3.3. Pooling

Pooling layers are non-linear functions, commonly maximum, minimum or average, applied to input vectors in order to perform a down-sampling on it [Collobert et al. 2011]. In the NLP context, pooling layers can be applied too. For this application, the most common way to use pooling layers is by applying a max-over-time pooling [Collobert et al. 2011], a global pooling function in the input vector, which in NLP tasks would be the output of an RNN. Finally, we stress that the application of pooling layers brings other benefits, such as reducing the size of input vectors and detecting invariant features.

4. Automatic Misinformation Detection with MIDeepBR

This section describes our approach, called MIDeepBR, a deep learning architecture for automatic misinformation classification in PT-BR WhatsApp Messages.

4.1. MIDeepBR Architecture

MIDeepBR makes use of BiLSTM, BERT Embeddings, Poolings, and Linear layers, as shown in Figure 1. Our approach combines these different tools and algorithms to improve the automatic misinformation detection performance. The entire MIDeepBR architecture is shown in Figure 1.

First, our textual data goes through a BERT [Devlin et al. 2018] layer, a type of transformer [Vaswani et al. 2017], a very popular Natural Language Processing technique. The BERT layers act as text embedding when using the last hidden state as a word vector. With this strategy, we take advantage of all BERT transformers and attentions tools leading to robust word's numeric representation.

We apply a Batch Normalization to the word vectors to improve the model's generalization, so now we can use them as input for a Bidirectional Long-Short Term Memory (BiLSTM) [Graves and Schmidhuber 2005] layer being able to capture the message context in forward and backward. Since there are messages of all types, we want to make sure that we can analyze the whole message in both directions to extract most of the information from them and completely understand their word sequence context. LSTM layers work very well with long sequences [Hochreiter and Schmidhuber 1997], so the model will be able to work with different text lengths.

After all word vectors of a message go through the BiLSTM, we again apply Batch Normalization on the BiLSTM outputs and use the Dropout [Srivastava et al. 2014] technique to avoid over-fitting by randomly turning off a portion of BiLSTM cells. We perform a Max Pooling and Average Pooling with the adjusted outputs, then concatenate them with the BiLSTM last hidden state output. These two pooling layers capture the most important (max pooling) outputs and the average value (average pooling) of the outputs. In other words, we can capture the most important words for the misinformation detection with max pooling and how all the words contribute to the detection with the average pooling. With this concatenation results, we feed it into linear layers to perform the classification.

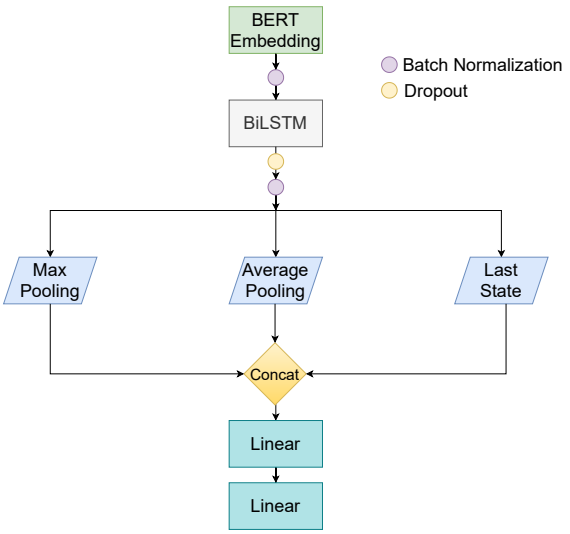

Figure 1. The MIDeepBR Architecture.

5. Experimental Evaluation

In this section, we describe the data set used in the experiments, called COVID-19.BR [Martins et al. 2021], and the adaptations that we performed on it. Besides, we present the algorithms and techniques used to build the automatic misinformation detection models and the performance metrics to evaluate them.

5.1. COVID-19.BR Data Set

An important aspect to consider while developing an appropriate method of automatic misinformation detection for WhatsApp messages in PT-BR is the necessity of a large-scale labelled data set. The COVID-19.BR [Martins et al. 2021] is a large-scale, labelled, anonymized, and public data set formed by WhatsApp messages in PT-BR about coronavirus pandemic, collected from public WhatsApp groups. COVID19-BR is inspired by [Silva et al. 2020] and the authors built it following the methodological guideline for building corpora of deceptive content of [Rubin et al. 2015].

COVID-19.BR contains messages from 236 open WhatsApp groups with at least 100 members. The data set has messages collected between April and June 2020. Alongside the messages, the data set other columns are date, hour, phone number, international phone code, if the user is Brazilian its state, word count, character count, and if the message contained media (audio, image, or video). Another feature of this data set is the definition of "viral messages" that are messages with more than five words that appear more than once in it. COVID-19.BR tackles users' privacy issues by anonymizing their names and cell phone numbers. Using a hash function to create a unique and anonymous identifier for each user using the cell phone number as input. It sets an alias for each group to achieve their anonymization. Since these groups are publicly available, this approach does not violate WhatsApp's privacy policy[7].

We revised the labels, removed the messages that have less than five words, messages not related to the coronavirus pandemics, messages containing only daily news summaries, and messages with only *url* as text content. The final corpus contains 2043 messages being 865 labelled as misinformation (label 1) and 1178 labelled as non-misinformation (label 0).

Table 1 presents 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 data set. We have a class ratio of 1.36, meaning that the data set is slightly imbalanced. Messages containing misinformation, on average, have more words, and their length varies more than the messages without misinformation, indicating that this type of message is disseminated in different writing styles. Number of shares has similar values in messages of both classes in our data set.

5.2. Experimental Environment

We reproduced the experiments from [Martins et al. 2021] to assess their performance on this revised data set by combining multiple text embedding techniques and classification algorithms. We added a new set of experiments using the MIDeepBR approach that uses deep learning techniques for text embedding and classification.

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

Table 1. Data set basic statistics.

Statistics	Non-misinformation	Misinformation
Count of unique messages	1178	865
Mean and std. dev. of number of tokens in messages	82.80 ± 57.59	169.72 ± 243.76
Minimum number of tokens	5	5
Median number of tokens	22	52
Maximum number of tokens	2210	1666
Mean and std. dev. of number of types in messages	57.59 ± 96.59	109.03 ± 133.15
Average size of words (in characters)	5.82	5.12
Type-token ratio	0.696	0.642
Mean and std. dev. of shares	2.02 ± 4.17	1.89 ± 2.76

5.2.1. Features and Algorithms

[Martins et al. 2021] explored vectors created with binary BoW and with the TF-IDF technique using different n-gram values, experimenting with unigrams, bigrams, and trigrams. Because of the lexical diversity of the corpus, the resulting vectors have large dimensions and sparsity. Even if that approach generates a larger vector space, the authors state that combinations of bigrams and trigrams can reveal distinguishable patterns present in messages with misinformation in the COVID-19.BR data set. So, combining 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) leads to a total of 12 different feature extraction scenarios.

For each scenario, we reproduced the experiments using the same nine machine learning classification techniques: 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 an RBF kernel [Prasetijo et al. 2017] (SVM), K-nearest neighbors (KNN), random forest (RF), gradient boosting (GB), and multilayer perceptron neural network (MLP). We used the Python scikit-learn [Pedregosa et al. 2011] module for all machine learning techniques. All the other machine learning techniques were used with default hyperparameters.

We trained MIDeepBR using the *adabelief* [Zhuang et al. 2020] optimizer, with learning rate of *0.00001* for 100 epochs. We used 650 hidden state features of the BiLSTM, a batch size of 128, and maximum length of messages of size 512. All messages with more than 512 tokens will be truncated due to BERT input size limitation. We developed the architecture using *pytorch* [Paszke et al. 2019] and the python module *transformers* [Wolf et al. 2019]. We used the *BERTimbau* [Souza et al. 2020] in our experiments, because it is a BERT model trained on PT-BR data.

5.2.2. Evaluation Metrics

We evaluated the performance of the experiments using the following metrics:

Table 2. Top 10 best combinations of classifiers and features extraction techniques using the corpus of the COVID-19.BR data set.

Rank	Experiment	Vocab.	FPR	PRE	REC	F1
1	MIDEEPBR-LEMMA	-	0.202	0.770	0.913	0.834
2	BOW-UNIGRAM-LEMMA-MLP	15014	0.149	0.799	0.800	0.799
3	BOW-TRIGRAM-LEMMA-MLP	189419	0.149	0.811	0.768	0.788
4	TFIDF-TRIGRAM-LEMMA-LSVM	189419	0.164	0.772	0.799	0.784
5	TFIDF-BIGRAM-LEMMA-LSVM	88501	0.161	0.780	0.791	0.784
6	TFIDF-UNIGRAM-LSVM	17666	0.156	0.794	0.773	0.783
7	TFIDF-BIGRAM-LSVM	104484	0.159	0.789	0.777	0.782
8	TFIDF-BIGRAM-LEMMA-MLP	88501	0.161	0.782	0.783	0.782
9	TFIDF-BIGRAM-MLP	104484	0.165	0.775	0.792	0.782
10	TFIDF-TRIGRAM-LSVM	243780	0.165	0.772	0.792	0.781

- False positive rate (FPR): the proportion of messages incorrectly classified as misinformation.
- Precision (PRE): the proportion of messages classified as misinformation and that truly belong to the misinformation class.
- Recall (REC): the proportion of misinformation correctly classified.
- F1-score (F1): the harmonic average between precision and recall.

Considering we are working with a binary classification task, where non-misinformation represents the negative class and misinformation the positive, these performance metrics are appropriate.

6. Results

For the sake of readability, we report only the top 10 best combinations of classifiers and text embedding. The results presented in the following tables are the metrics' mean after 5 rounds of k-fold cross-validation, except for the MIDeepBR approach due to computational limitations.

Table 2 summarizes the results for the experiments we run considering the full corpus of the COVID-19.BR data set [Martins et al. 2021]. Analyzing the F1 results, we can observe that the MIDeepBR approach outperforms by 3.5%, the best result using classic machine learning models and text embedding. Although the FPR of the MIDeepBR is 5.1% higher than the best classic model, the model's bias to predict messages as misinformation. MIDeepBR's PRE is lower due to the model bias since misinformation is the minority class in the data set. Despite the high FPR metric value, MIDeepBR can correctly predict misinformation messages better than any classic approach, with a REC of 0.913. In this problem of misinformation detection, it is better to choose models with better REC than with less FPR, but also, we can not forget to evaluate the F1 values to make sure the predictions are balanced. Removal of stop words and lemmatization are present in the best-performing models. The best classic model was the MLP trained with BoW as text embedding, unigram, removing stop words and performing lemmatization. LSVM also performed very well in these experiments.

In [Martins et al. 2021], the authors performed another experiment using only

Table 3. Top 10 best combinations of classifiers and features extraction using only the long messages of the COVID-19.BR data set.

Rank	Experiment	Vocab.	FPR	PRE	REC	F1
1	BOW-BIGRAM-NB	93589	0.184	0.828	0.913	0.866
2	BOW-TRIGRAM-NB	216068	0.188	0.829	0.908	0.864
3	TFIDF-TRIGRAM-LEMMA-MLP	167089	0.199	0.825	0.915	0.862
4	TFIDF-TRIGRAM-MLP	216068	0.197	0.818	0.910	0.860
5	BOW-TRIGRAM-LEMMA-NB	167089	0.200	0.817	0.915	0.860
6	BOW-UNIGRAM-NB	16335	0.188	0.834	0.890	0.859
7	TFIDF-BIGRAM-LEMMA-MLP	78838	0.196	0.830	0.899	0.859
8	BOW-BIGRAM-LEMMA-NB	78838	0.201	0.815	0.913	0.858
9	TFIDF-UNIGRAM-LEMMA-MLP	13955	0.211	0.799	0.928	0.857
10	TFIDF-UNIGRAM-MLP	16335	0.218	0.789	0.939	0.856

messages containing 50 or more words in the COVID-19.BR data set. We also reproduced this experiment. The reason behind this new set of experiments is to analyze if the text length influences the automatic misinformation detection models' performance. The resulting subset has 822 messages being 446 containing misinformation and 376 not containing misinformation. Table 3 shows the results for this second scenario. We can observe that in terms of F1, the performance increased in this scenario. The best model achieved an F1 of 0.866 using BoW, unigram, and NB as the combination of embedding and classifier, the same model combination reported by [Martins et al. 2021]. This model also achieved the lowest FPR, 0.184, among the candidates. Only NB and MLP appeared in the top 10 combinations.

It is important to highlight that MIDeepBR did not figure in the top 10 best models (Table 3), appearing at rank 70 in our experiments, achieving an F1 of 0.826. However, this result is due to the BERT layer input size limitation of 512 tokens.

We can observe from all the results that MIDeepBR performs well in the general case, but its performance is not improved when trained with only the long texts. On the other hand, classic machine learning approaches perform well when trained with only long texts. This happens because of the BERT layer input size limitation, which causes MIDeepBR not to train with all the information available in those messages, while the classic approaches can deal with messages of any size. Because it is a more complex model that uses deep learning and advanced NLP techniques, MIDeepBR's training time is 756 minutes using GPU to train it, while the maximum training time of classic approaches is 22 minutes using CPU to train them. All the experiments and the COVID-19.BR data set are available at our public repository[8].

7. Conclusion

In these days of pandemics, the automatic misinformation detection (MID) about COVID-19 in PT-BR WhatsApp messages is a crucial challenge. The early detection of misinformation can prevent its spread, thus reducing its damage. This paper presented MIDeepBR, a new approach based on BiLSTM neural networks, pooling operations, and attention

[8]https://gitlab.com/jmmonteiro/misinformation_covid19_mideepbr

mechanisms. The results we reached experimenting with our approach indicated that it can automatically detect misinformation in PT-BR WhatsApp messages as soon as it is available at the Digital Lighthouse [de Sá et al. 2021] platform.

In the experiments performed on the COVID-19.BR data set, the MIDeepBR approach achieved an F1 score of 0.834, while in previous works, the best results achieved an F1 score of 0.778. Thus, MIDeepBR outperforms the previous works. However, the MIDeepBR approach can not perform very well when considering only long messages due to the BERT layer input size limitation.

As future works, we want to apply eXplainable Artificial Intelligence (XAI) tools to our models to understand better how these models treat the messages. We also ought to perform more qualitative analysis. To increase the performance of the classifiers, we want to assess the usage of the *Longformer* [Beltagy et al. 2020] as the text embedding layer to avoid the input size limitation from BERT.

References

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

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.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

de Sá, I. C., Monteiro, J. M., da Silva, J. W. F., Medeiros, L. M., Mourão, P. J. C., and da Cunha, L. C. C. (2021). Digital lighthouse: A platform for monitoring public groups in whatsapp. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, pages 297–304. SCITEPRESS.

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.

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.

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.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

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

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.

Maakoul, O., Boucht, S., El Hachimi, K., and Azzouzi, S. (2020). Towards evaluating the covid'19 related fake news problem: Case of morocco. In *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, pages 1–6.

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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.

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.

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.

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.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

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.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S., Dvornek, N. C., Papademetris, X., and Duncan, J. S. (2020). Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *CoRR*, abs/2010.07468.