ISWARA at WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets using BERT and FastText Embeddings

Wava Carissa Putri*[1], Rani Aulia Hidayat*[1],
Isnaini Nurul Khasanah*[1], Rahmad Mahendra[2]
Faculty of Computer Science, Universitas Indonesia
[1]{wava.carissa01, rani.auila,isnaini.nurul91}@ui.ac.id
[2]rahmad.mahendra@cs.ui.ac.id

Abstract

This paper presents Iswara's participation in the WNUT-2020 Task 2 "Identification of Informative COVID-19 English Tweets using BERT and FastText Embeddings", which tries to classify whether a certain tweet is considered informative or not. We proposed a method that utilizes word embeddings and using word occurrence related to the topic for this task. We compare several models to get the best performance. Results show that pairing BERT with word occurrences outperforms fastText with F1-Score, precision, recall, and accuracy on test data of 76%, 81%, 72%, and 79%, respectively.

1 Introduction

Twitter is known for being one of the major platforms during disasters (Ashktorab et al., 2014) since it is accessible for everyone and could give real-time information about the disaster (Vieweg, 2010). The outbreak of COVID-19 causes a large flow of information about the pandemic within the site. The increase in the number of tweets raised concern about the relevancy of the tweets to the COVID-19 itself. Identifying relevant tweets manually form Twitter is costly and needs significant human efforts (Nguyen et al., 2020).

WNUT-2020 has organized a shared task which focuses on identifying whether an English tweet related to COVID-19 is informative or not. Tweets containing information about recovered, suspected, confirmed, and death cases of COVID-19 is defined as an informative tweet. This paper aims to present the approaches we developed as part of our participation in WNUT-2020 Task 2. In this research, we proposed two approaches to identify informative COVID-19 tweets. The first approach is based on the BERT model, whereas the second approach is based on the FastText model. The first approach is experimented on two classification models such as Logistic Regression (LR) and Support Vector Machine (SVM). Word occurrence also used as an additional feature during the experiment. The second approach uses a built-in classifier from the FastText library.

The rest of the paper is organized as follows. In section 2, we explain the methodology we developed. Section 3 describes the results analysis about the experiments. Finally, section 4 concludes the paper and lists the future works.

2 Methodology

We first analyze the training data provided by the task organizers. It consisted of 7,000 tweets, with 3,697 tweets labeled as uninformative and 3,303 tweets as informative. The validation data consisted of 1000 tweets, with 528 uninformative tweets and 472 informative tweets. Most of the tweets were written in English. However, we found that there are several words that were written in non-ASCII characters.

The rise of word embedding led us to use pre-trained word embedding models such as BERT and fastText in our experiments. To process with the pre-trained model, we had to preprocess the tweets beforehand. We tried to remove URLs, Non-ASCII characters, tokens such as 'RT', '@', '&','<', '>' and extra space. All of the tweets were changed into lowercase letters.

[*] These authors contributed equally to this work

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 491–494
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics

In this experiment, we used pre-trained word embedding models to generate a word embedding representation of each tweet. Each representation is fed into a classifier to identify which class the tweet belongs to. BERT (Devlin et al., 2018) is used to create a representation of the tweets. Meanwhile, fastText (Joulin et al.,2017) is used directly for the tweet classification after modifying the label by adding the term “__label__” followed by the actual label.

Prior research found that BERT (Müller et al., 2020; Roitero et al., 2020) and fastText (Stein et al., 2019; Jha and Mamidi, 2017; Alessa et al., 2018) are useful for tweet classification tasks. There are a variety of pre-trained models available for BERT. However, we decided to use DistilBERT (Sanh et al., 2019) since it has a similar performance with other pre-trained models, but it has a much smaller representation model.

In addition, we also used word occurrences and number existence of the preprocessed tweets as the features that we used to classify the tweets. We combined the word occurrences with the word embedding representation. The occurrences of words ‘corona’ and ‘covid’ are being used as features to represent that a tweet is related to COVID-19. As Nguyen et al. (2020) stated, a tweet is considered as informative if it contains any information about recovered, suspected, confirmed, and death cases. Hence, the occurrences of words ‘recover’, ‘suspect’, ‘confirm’, ‘death’, and ‘case’ are used. Moreover, the number of cases is sometimes mentioned in the tweets. We covered this by using the existence of numerical value as the feature. The number existence feature will be true if a tweet contains a numerical value, and false if there is no numerical value in the tweet.

There are several classifiers that we used in this study to identify the informative tweets. Those classifiers are Logistic Regression, Support Vector Machine, and Multinomial Logistic Regression (built-in classifier in fastText library). Furthermore, we compared the performance of each classifier to get the best model.

Logistic Regression (LR) is a machine learning method that calculates the result by considering each feature’s weight. It has been applied to classify tweets into certain topics (S.T. et al., 2016). After tuning the parameters, we decided to train our LR model with a value of 5.263252631578947 for C and a maximum iteration of 10,000.

Support Vector Machine has been proven to give acceptable performance for tweets classification (Kurniawan et al., 2016). This classifier is well known for its ability to map nonlinear data into a higher dimensional space using the kernel trick. In this study, we focused on implementing the RBF kernel. Parameters C and gamma have a significant role in SVM with RBF kernels. The value of parameters C and gamma that we used to train the SVM model are 10 and 0.01, respectively. Those parameters are obtained after the tuning process.

Joulin et al. (2017) showed that fastText gives on par accuracy compared to deep learning classifiers but with faster training and validation process. As mentioned earlier, we labeled each tweet with the text “__label__” followed by the actual label (informative or uninformative). Then, we train the model on the preprocessed train data using the fastText library. This library will extract fastText embeddings and process it with the built-in classifier, Multinomial Logistic Regression (MLR). To find the best hyperparameters, we tried automatic and manual hyperparameter optimization. Based on the evaluation score on the validation data, we find better results using hyperparameters that we find manually. The value of hyperparameters learning rate, maximum length of word n-gram, epoch, minimum number of word occurrences, number of buckets, and loss function used in this study are 0.075, 2, 150, 6, 200000, and ‘hs’, respectively.

3 Results and Analysis

There are several scenarios of features in this study. The first scenario is using BERT as the feature. In the second scenario, we combined word occurrences (WO) with BERT as the features. We employed the first and second scenarios to LR and SVM with parameters we defined in the previous section. In the next scenario, we used the preprocessed tweets as the feature for fastText to generate fastText embbedings (FT).

Since the task is focused on informative tweets' performance, the evaluation metrics we compare in our study are F1-Score, precision, recall, and accuracy of the informative labeled tweets. Before submitting the final model to be evaluated using test data on WNUT-2020 system, we evaluated the model by the validation data. Figure 1 shows all experiments that have been done in this study.

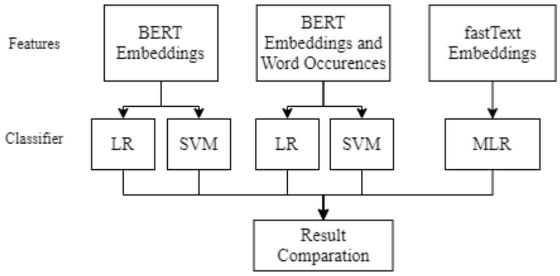

Figure 1. Experiment Scenarios

Table 1 shows the result of each experiment evaluated on validation data. The best F1-score is 0.81 and it was obtained by two models. The first model used preprocessed tweets with fastText, which resulted in precision, recall, and accuracy of 0.85, 0.77, and 0.83, respectively. The next model used BERT and word occurrences as features with SVM as classifier with precision, recall, and accuracy of 0.83, 0.78, and 0.82, respectively.

Features	Classifier	F1-Score	Precision	Recall	Accuracy
FT	MLR	0.81	0.85	0.77	0.83
BERT	LR	0.79	0.81	0.77	0.81
BERT	SVM	0.79	0.82	0.76	0.81
BERT & WO	LR	0.79	0.82	0.77	0.81
BERT & WO	SVM	0.81	0.83	0.78	0.82

Table 1. Result on Validation Data

From the experiments we conducted, the performance of BERT increased when combined with word occurrences as the additional features whether it is employed to LR or SVM. It is shown that using COVID-19 related word occurrences as features successfully manages the model to identify informative tweets better.

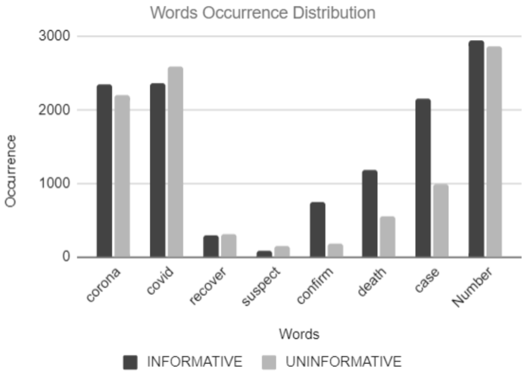

Figure 2. Word Occurrences Distribution

Figure 2 shows that there is not much difference in the occurrence of words 'corona', 'covid', 'recover' and numerical value in informative and uninformative tweets. Meanwhile, there is a high difference in the occurrence of words 'confirm', 'death', and 'case' between informative and uninformative tweets. It means that the word occurrences help the model to classify the tweets.

Because of the limited number of submissions on WNUT-20202 system on Task 2, we chose the best two models based on the experiments result. These two models were officially evaluated on test data in the system provided by WNUT-2020.

Table 2 presents the performance of each model. It is shown that the performance of the model, which used BERT and word occurrences with SVM, outperforms the model that used fastText with MLR. The best model achieved an F1-Score of 0.763, precision of 0.807, recall of 0.723, and accuracy of 0.788.

Features	Classifier	F1-Score	Precision	Recall	Accuracy
BERT & WO	SVM	0.763	0.807	0.724	0.788
FT	MLR	0.757	0.791	0.725	0.780

Table 2. Result of Test Data on WNUT-2020 System.

We found that fastText performs better when evaluated on the validation data. However, the combination BERT and word occurrences shows better results when evaluated on the test data. One reason may be that since BERT created subwords rather than n-grams in their approach, it could help the model to handle out of vocabulary words in the test data better.

4 Conclusion

This paper presents our approach for WNUT-2020 Task 2 to identify whether a tweet related to COVID-19 is informative or not. The proposed method for this shared task is combining word embedding with word occurrences as features. We compared several scenarios of features with different classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), and Multinomial Logistic Regression (MLR), built-in classifier in fastText library, on validation data.

The validation data result showed that there are two models that give the highest F1-Score of 0.81. The first model used the fastText embeddings with built-in MLR has precision, recall, and accuracy of 0.85, 0.77, and 0.83, respectively. The second model used the BERT embeddings and word occurrences as features with SVM as the classifier has precision, recall, and accuracy of 0.83, 0.78, and 0.82, respectively. These two models are evaluated on the WNUT-2020 system with test data. On the test data, the model that used BERT embeddings and word occurrences with SVM outperforms the model which used fastText embeddings with MLR. It achieved an F1-Score of 0.763, precision of 0.807, recall of 0.723, and accuracy of 0.788.

The use of richer features such as Named Entity Recognition to identify location mentioned in a tweet and the use of Part of Speech Tagging can be done as the future works. Moreover, the implementation of advanced algorithms such as deep learning might also increase the performance of the model to identify informative tweets.

References

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. 2014 . Tweedr: Mining twitter to inform disaster response. In *Information Systems for Crisis Response and Management (ISCRAM)*, pages 269-272.

Vieweg, S. 2010. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work*, pages 515-516.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop Noisy User-generated Text*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 2: Short Papers). Association for Computational Linguistics, pages 427-431.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. *COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter*. arXiv preprint arXiv:2005.07503.

Roger Alan Stein, Patricia A. Jaques, and João Francisco Valiati. 2019. *An analysis of hierarchical text classification using word embeddings*. Information Sciences, 471, pages 216-232.

Aksitha Jha, and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*. Association for Computational Linguistics, pages 7-16.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

Indra S.T., Liza Wikarsa, and Rinaldo Turang. 2016. Using Logistic Regression Method to Classify Tweets into the Selected Topics. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 385-390.

Dwi Aji Kurniawan, Sunu Wibrama, and Nur Akhmad Setiawan. 2016. Real-time Traffic Classification with Twitter Data Mining. *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1-5.

Kevin Roitero, Cristian Bozzato, Vincenzo Della Mea, Steffano Mizzaro, and Giuseppe Serra. 2020. Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT. In *Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH) 2020.*

Ali Alessa, Miad Faezipour, and Zakhriya Alhassan. 2018. Text classification of flu-related tweets using fasttext with sentiment and keyword features. *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 366-367.