CSECU-DSG at WNUT-2020 Task 2: Exploiting Ensemble of Transfer Learning and Hand-crafted Features for Identification of Informative COVID-19 English Tweets

Fareen Tasneem, Jannatun Naim, Radiathun Tasnia,
Tashin Hossain, and Abu Nowshed Chy
Department of Computer Science & Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{fareen.tasneem,jannatun.naim.cu,radia.tasnia.cu,
tashin.hossain.cu}@gmail.com and nowshed@cu.ac.bd

Abstract

COVID-19 pandemic has become the trending topic on twitter and people are interested in sharing diverse information ranging from new cases, healthcare guidelines, medicine, and vaccine news. Such information assists the people to be updated about the situation as well as beneficial for public safety personnel for decision making. However, the informal nature of twitter makes it challenging to refine the informative tweets from the huge tweet streams. To address these challenges WNUT-2020 introduced a shared task focusing on COVID-19 related informative tweet identification. In this paper, we describe our participation in this task. We propose a neural model that adopts the strength of transfer learning and hand-crafted features in a unified architecture. To extract the transfer learning features, we utilize the state-of-the-art pre-trained sentence embedding model BERT, RoBERTa, and InferSent, whereas various twitter characteristics are exploited to extract the hand-crafted features. Next, various feature combinations are utilized to train a set of multilayer perceptron (MLP) as the base-classifier. Finally, a majority voting based fusion approach is employed to determine the informative tweets. Our approach achieved competitive performance and outperformed the baseline by 7% (approx.).

1 Introduction

Twitter is one of the most prominent microblogging platforms that provides a convenient way of sharing opinions and broadcasting news and information briefly among the mass community. People often prefer using twitter due to its real-time feature because compared to others, it contains most of the notable or official information. Hence, information shared on this platform is very helpful during emergency situations. That is why, in the ongoing COVID-19 pandemic, people are interested in seeking informative tweets related to COVID-19. An informative tweet may contain the updated information of COVID-19, new cases or death information, medicine or vaccine news, and updated guidelines from diverse sources.

Tweet#1: BREAKING: 21 people on Grand Princess cruise ship docked off the California coast tested positive for coronavirus ... **Label:** INFORMATIVE
Tweet#2: The WHO is being haunted by an old tweet saying that China found no human transmission ..... **Label:** UNINFORMATIVE

Table 1: Example of sample tweets with labels.

For instance, in Table 1 Tweet#1 is an informative tweet since it contains the updated news of coronavirus affected people on a cruise ship. Besides, Tweet#2 doesn't contain any valuable news or information, instead it provokes annoyance among the people who are keen to know the updates of the COVID-19 situation. Therefore, refining the informative tweets from the real-time tweet stream is a formidable task.

Classifying informative tweets on COVID-19 is an emerging concept. Nevertheless, there are some prior works on identifying informative tweets during crises or disasters. For example, a CNN based approach proposed by (Caragea et al., 2016) for informative tweets identification in disaster. Later, (Fadaei et al., 2018) and (Neppalli et al., 2018) employed CNN with GloVe and Word2Vec features, respectively in the same context. More recently, a work regarding misinformation detection on COVID-19 tweets (Hossain et al., 2020) has also been done.

In this paper, we present our proposed systems submitted to the W-NUT 2020 shared task 2 (Nguyen et al., 2020). The primary goal of the task is to identify informative tweets related to

The first four authors have equal contributions.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 394–398
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics

COVID-19. We exploit the transfer learning features from BERT, RoBERTa, and InferSent along with the n-gram and other hand-crafted features in a unified neural model.

The rest of the contents are structured as follows: we describe our proposed framework in Section 2 whereas Section 3 includes the experimental details and performance analysis. We concluded this paper with plausible future notions in Section 4.

2 Proposed Framework

In this section, we describe our proposed framework. Our target is to identify the informative tweet based on its context. The overview of our proposed framework is presented in Figure 1.

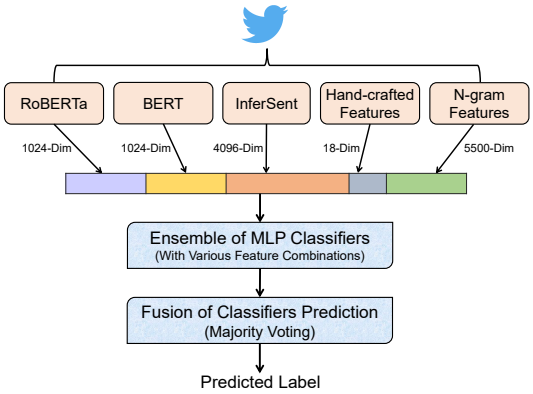

Figure 1: Proposed Framework.

Given a tweet, we explore various techniques to encode the tweet into effective feature vectors. We utilize three pre-trained embedding models including RoBERTa, BERT, and InferSent to extract the effective transfer learning features. Besides, we extract the n-gram and other hand-crafted features. The concatenated features are then passed to a multilayer perceptron (MLP) network for the training. We explore different feature combinations to train and extract the prediction from multiple MLP classifiers. Finally, a majority voting based fusion scheme exploits all the predicted labels and determine the final one.

2.1 Tweet Encoding

BERT: BERT (Devlin et al., 2019) stands for bidirectional encoder representations from transformers, which is a new method of pre-training sentence representations. It captures the context for every existence of given tweets. We employ the BERT-Large, Uncased model to encode each tweet into a 1024-dimensional feature vector included with 24 layers, 16 heads, and 340M parameters.

RoBERTa: RoBERTa (Liu et al., 2019) is considered as an extended version of BERT which is called a robustly optimized BERT pretraining approach. It emphasis on the key hyper-parameters choices and ignoring the next sentence prediction (NSP) objective. It also uses a larger mini-batches and learning rates while training the model. We conduct the RoBERTa-Large pre-trained model based on BERT-Large architecture for encoding each tweet into a 1024-dimensional feature vector.

InferSent: InferSent (Conneau et al., 2017) indicates a sentence encoding approach that is based on natural language interference data for semantically highlighting the sentence through vectorized representations. We exploit the InferSent pre-trained sentence encoding model which is trained with fastText and embedded each tweet into a 4096-dimensional feature vector.

N-gram Bag-of-Words (Bow) Features: The bag-of-words (BoW) representation models a tweet based on its word occurrence statistics. To extract the n-gram features, we utilize maximum features length into a 5500-dimensional feature vector based on word occurrence and use the n-gram range of (1, 8) for capturing the diverse types of features.

To extract the effective n-gram features, we employ the data preprocessing that reduces the effect of noise. We perform the stemming using the Snowball stemmer and remove the accented and special characters from tweets. The tweets may contain some non-standard words (e.g. 2day, suspct). To normalize such noisy words we follow a similar approach used by (Chy et al., 2017), where they utilized two lexical normalization dictionaries (2012; 2012) to address this problem. Besides hashtags may contain important information and segmenting the hashtag (e.g. #CovidPandemic to covid pandemic) might be beneficial to distill the content of the tweets. We utilize a tool provided by Baziotis et al. (2017) to segment the hashtag. Besides, we also demojize (emoji to text) the emoji using a tool[2], expand the contradictions of words, and convert the number into words using PyPi inflection[3].

Other Hand-crafted Features: We extract a set of 18 handcrafted features that comprise of textual features, tweet-specific features, emoticon-based features, POS features, sentiment-based features, and COVID-19 related features. A COVID-19 related informative tweet may contain the news

[2]https://github.com/NeelShah18/emot/
[3]https://pypi.org/project/inflection/

Sl.	Feature Definition
F1	*Average Word Length:* Average of characters per word in a tweet.
F2	*Tokenized Words Count:* Number of tokenized words in a tweet (Wang et al., 2019).
F3	*Symbol Count:* Number of symbols present in a tweet (Wang et al., 2019).
F4	*Capital Word Ratio:* The ratio of capital words to total words in a tweet (Wang et al., 2019).
F5	*Digit Ratio:* The ratio of digits to total characters in a tweet (Wang et al., 2019).
F6	*Digits Count:* Number of digits in a tweet (Wang et al., 2019).
F7	*URL Count:* Number of URLs present in a tweet (Pamungkas and Patti, 2018).
F8	*Retweet Count:* Number of retweets present in a tweet (Siddiqua et al., 2016).
F9	*Hashtag Count:* Number of hashtags present in a tweet (Pamungkas and Patti, 2018).
F10	*Number of Emoticons:* Number of emoticons in a tweet (Hogenboom et al., 2015).
F11	*Presence of Emoticon:* Checks if a tweet contains emoticons (Hogenboom et al., 2015).
F12	*Adverb Count:* Number of adverbs in a tweet (Siddiqua et al., 2016).
F13	*Polarity Score:* Determines the sentiment of the tweet (Swanberg et al., 2018) using TextBlob (2014).
F14	*Positive Sentiment Presence:* Checks if a tweet contains positive sentiment using SentiStrength (2012).
F15	*Slang Word Presence:* Checks if slang words are present in a tweet or not using slang word lexicon (2012).
F16	*Special Verb Count:* Number of special verbs (i.e. related to COVID context) in a tweet.
F17	*COVID Cases Presence:* Checks if a tweet contains number of COVID cases or deaths.
F18	*Keywords Count:* Number of COVID related keywords(e.g. tested, quarantine, etc.) in a tweet.
Total	18 Features

Table 2: List of other hand-crafted features used in this work.

about new cases, deaths, and recoveries. We exploit these characteristics and extract the COVID cases presence feature, where we check the presence of numerical value along with some specific keywords (e.g. deaths, affected cases). We have also inspected the presence of some COVID-19 specific keywords and verbs that are more frequent in informative tweets. We built two lexicons and extract the special verb count and keywords count features, accordingly. The features definitions are presented in Table 2.

2.2 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) (Windeatt, 2006) is a supervised learning method where backpropagation is used for training. It can be separated from the linear perceptron because of its multi-layer structure and non-linear process. MLP classifier includes various constructive parameters including input layer, hidden layers, output layer, iterations, learning rates, and activation functions. We utilize our MLP classifier with two hidden layers with 500 and 2000 neurons, relu activation function, and set alpha as 0.003. Besides, we use the maximum iterations of 1000 with a random state 1 and a tolerance rate of 0.0001 in our proposed system.

2.3 Fusion of Classifiers Predictions

We empirically exploit various feature combinations in our classifier settings and obtain the prediction from N-number of classification systems. The intuition behind this process is to distill different types of contextual information in different settings. Next, we use the majority voting count based scheme to obtain the final prediction label. The core idea behind this type of classification is that the final output class is selected based on the highest number of votes (2010). In prior research, this classification scheme was used to overcome the limitations of a single classifier (2016).

3 Experiments and Evaluations

3.1 Dataset Description

For evaluating our informative tweet identification system, we made use of the COVID-19 Tweet dataset (2020) having 10K COVID-19 related English Tweets. The training set and valid set consists of 7K and 1K tweets with the valance portion of Informative and Uninformative tweets. Along with that, the test dataset consists of 944 Informative and 1056 Uninformative tweets. The detailed statistics were presented in Table 3.

Category	#Train	#Dev	#Test	#Total
Informative	3303	472	944	4719
Uninformative	3697	528	1056	5281
Total	7000	1000	2000	10000

Table 3: The statistics of the dataset.

Following the benchmark of WNUT-2020 Task 2 (2020), we used the accuracy, precision, recall, and F1 score as the evaluation measures, where the F1 score is considered as the primary metric.

3.2 Experimental Setup and Results Analysis

We now describe the settings of our submitted system to the WNUT-2020 Task 2 and analyze the informative tweet identification performances. In our CSECU-DSG#1 system, we have exploited various feature combinations to obtain the predictions from N=7 different classification systems. The feature settings are described in Table 4. Based on the prediction from these 7 systems, a majority voting based method is employed to determine the final label. Besides, we utilize the feature settings $S2$ in Table 4 and the corresponding classifier settings in our CSECU-DSG#2 system.

Sl.	Various Feature Settings Used in CSECU-DSG#1
S1	BERT, N-gram, Other HCF (F9,F7,F8,F10,F16)
S2	RoBERTa ,BERT, N-gram, Other HCF (F9,F7,F8,F11,F10,F16,F12,F14,F15,F18)
S3	RoBERTa ,BERT, N-gram, Other HCF (F1,F3)
S4	RoBERTa ,BERT, N-gram, Other HCF (F7,F8,F9,F10,F16)
S5	InferSent, RoBERTa, BERT, N-gram, Other HCF (F7,F8,F9,F10,F16)
S6	RoBERTa, BERT, N-gram, Other HCF (F14)
S7	RoBERTa, BERT, N-gram, Other HCF (F2,F4,F5,F6,F13,F16,F17,F18)

Table 4: Feature settings of various classifiers.

Now, we compare the performance of our submitted systems against the other participants' systems and the baseline. The organizers used the FastText classifier as the baseline system. The comparative performance are presented in Table 5.

Team_Name	F1 score	Precision	Recall	Accuracy
NutCracker	0.9096	0.9135	0.9057	0.9150
Husky	0.8992	0.8959	0.9025	0.9045
CSECU-DSG#1	0.8198	0.8155	0.8242	0.8290
CSECU-DSG#2	0.8156	0.8134	0.8177	0.8255
IIITBH	0.7979	0.7991	0.7966	0.8095
NLPRL	0.7854	0.8335	0.7426	0.8085
Baseline	0.7503	0.7730	0.7288	0.7710

Table 5: Comparative performance analysis (2020).

It shows that both of our systems achieved competitive performances and obtained nearly similar kinds of performances. This deduces that the classifiers ensemble (CSECU-DSG#1) based on different feature combinations (change mostly hand-crafted features) might not fit for the task since it did not achieve the substantial performance improvement over the single classifier (CSECU-DSG#2) on the test set. Besides, combining a rich set of features without employing effective feature selection techniques and learning models hampered the performance of our model. However, our CSECU-DSG#1 system lacks by $\approx 9\%$ from the top-performing system NutCracker but surpassed the FastText baseline by $\approx 7\%$.

3.3 Discussion

We conduct the feature ablation study to evaluate the effectiveness of different types of features including pre-trained embedding features, n-gram features, and other hand-crafted features. To do this, we utilized our proposed CSECU-DSG#2 based on the validation set. The results are shown in Table 6.

Method	F1 score	Precision	Recall	Accuracy
CSECU-DSG#2	0.8598	0.8553	0.8644	0.8670
Feature Ablation Study				
−RoBERTa	0.8145	0.7587	0.8792	0.8110
−BERT	0.8237	0.8253	0.8962	0.8190
−N-gram	0.8245	0.7828	0.8707	0.8250
−Other HCF	0.8474	0.8253	0.8708	0.8520

Table 6: Feature ablation study on the validation set.

It showed that RoBERTa has the highest impact on the system's performance and the primary metric F1 score decreased by 4.53% while removing this feature. Besides, the F1 score decreased by 3.61%, 3.53%, and 1.24% when ablating the BERT, N-gram, and other hand-crafted features, respectively. This deduced the contribution of these features for informative tweet identification.

4 Conclusion and Future Directions

In this paper, we have explored various transfer learning features along with a rich set of hand-crafted features in an MLP based unified neural framework to identify the COVID-19 related informative tweets. We analyzed the effect of each feature type on the classification performances. Our systems achieved competitive performances among the participants' systems.

In the future, we have a plan to exploit the vast amount of COVID-19 related informative tweets to train and downstream fine-tune the various transfer learning models for extracting effective tweet representations. We also aim to inspect more proficient hand-crafted features for identifying COVID-19 related informative tweets.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.

Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. 2016. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 137–147.

Abu Nowshed Chy, Md Zia Ullah, and Masaki Aono. 2017. Microblog retrieval using ensemble of feature sets through supervised feature selection. *IEICE TRANSACTIONS on Information and Systems*, 100(4):793–806.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 4171–4186.

Noushin Fadaei, Chanjong Im, Sandip Modha, and Thomas Mandl. 2018. Daiict-hildesheim@ information retrieval from microblogs during disasters (irmidis 2018). In *Forum for Information Retrieval Evaluation (FIRE) (Working Notes)*, pages 15–17.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 421–432. Association for Computational Linguistics (ACL).

Alexander Hogenboom, Danella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *Journal of Web Engineering*, 14(1-2):22–40.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sameer Singh, and Sean Young. 2020. Detecting covid-19 misinformation on social media. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*.

Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers-Volume 1*, pages 1035–1044. Association for Computational Linguistics (ACL).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: simplified text processing*, 3.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *ISCRAM*.

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 on Noisy User-generated Text*.

Endang Wahyu Pamungkas and Viviana Patti. 2018. # nondicevosulserio at semeval-2018 task 3: Exploiting emojis and affective content for irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 649–654.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Umme Aymun Siddiqua, Tanveer Ahsan, and Abu Nowshed Chy. 2016. Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 304–309. IEEE.

Kevin Swanberg, Madiha Mirza, Ted Pedersen, and Zhenduo Wang. 2018. Alanis at semeval-2018 task 3: A feature engineering approach to irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 507–511.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology (JASIST)*, 63(1):163–173.

Junpei Zhou Xinyu Wang, Po-yao Huang, and Alexander Hauptmann. 2019. Cmu-informedia at trec 2019 incident streams track. In *Proceedings of the 28th Text REtrieval Conference (TREC)*. NIST.

Terry Windeatt. 2006. Accuracy/diversity and ensemble mlp classifier design. *IEEE Transactions on Neural Networks*, 17(5):1194–1211.