DSC-IIT ISM at WNUT-2020 Task 2: Detection of COVID-19 informative tweets using RoBERTa

Sirigireddy Dhanalaxmi, Rohit Agarwal* and Aman Sinha*

Indian Institute of Technology (Indian School of Mines) Dhanbad, India
{sirigireddydhanalaxmi, agarwal.102497, amansinha091}@gmail.com

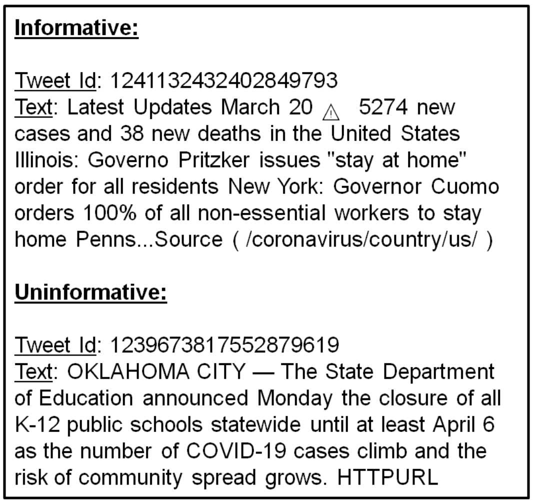
Abstract

Social media such as Twitter is a hotspot of user-generated information. In this ongoing Covid-19 pandemic, there has been an abundance of data on social media which can be classified as informative and uninformative content. In this paper, we present our work to detect informative Covid-19 English tweets using RoBERTa model as a part of the W-NUT workshop 2020. We show the efficacy of our model on a public dataset with an F1-score of 0.89 on the validation dataset and 0.87 on the leaderboard.

1 Introduction

Text analysis of social media data gives broader insights into various topics discussed among people. Twitter is a social media platform where people interact through short texts. This paper constitutes our work for the Shared Task 2 of the 6th Workshop on Noisy User-generated Text (W-NUT) (Nguyen et al., 2020) where we need to classify the Covid-19 English tweets as informative or uninformative. In the context of this shared task, a tweet is considered informative if it is about recovered, suspected, confirmed, and death cases and location or travel history of the cases, and all the other tweets fall into the category of uninformative class. Figure 1 shows an example of both informative and uninformative tweets.

We applied various machine learning models such as logistic regression, Naive Bayes, random forest classifier, support vector machine (SVM), and multi-layer perceptron (MLP). We have also used several state-of-the-art architectures like BERT, DistilBERT, RoBERTa, and ALBERT for detecting informative tweets. We provide a comparative study of all these models and found the RoBERTa model to perform best among all the models.



Informative:

Tweet Id: 1241132432402849793
Text: Latest Updates March 20 ⚠ 5274 new cases and 38 new deaths in the United States Illinois: Governo Pritzker issues "stay at home" order for all residents New York: Governor Cuomo orders 100% of all non-essential workers to stay home Penns...Source (/coronavirus/country/us/)

Uninformative:

Tweet Id: 1239673817552879619
Text: OKLAHOMA CITY — The State Department of Education announced Monday the closure of all K-12 public schools statewide until at least April 6 as the number of COVID-19 cases climb and the risk of community spread grows. HTTPURL

Figure 1: An example of informative and uninformative tweet.

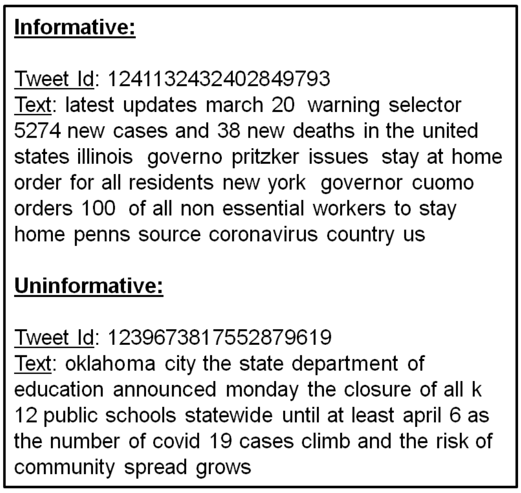
This paper's outline is as follows: Section 2 discusses the previous works related to our paper. Section 3 describes the dataset and the data preprocessing steps. Section 4 describes our methods and section 5 discusses the implementation details of our approaches. Section 6 contains the analysis of the results, which is followed by the conclusion in section 7.

2 Related Work

Detection of useful-crisis-related content has been pivoting around Twitter due to its interactive media via microtexts (Martinez-Rojas et al., 2018). Continuous Bag-of-Words (CBoW) based approach has been used for text classification (Sriram et al., 2010). Castillo et al. (2011) proposes the use of different user-based features representing messages and tweet propagation for classifying tweet credibility.

* Equal contribution.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 409–413
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics



Informative:

Tweet Id: 1241132432402849793
Text: latest updates march 20 warning selector 5274 new cases and 38 new deaths in the united states illinois governo pritzker issues stay at home order for all residents new york governor cuomo orders 100 of all non essential workers to stay home penns source coronavirus country us

Uninformative:

Tweet Id: 1239673817552879619
Text: oklahoma city the state department of education announced monday the closure of all k 12 public schools statewide until at least april 6 as the number of covid 19 cases climb and the risk of community spread grows

Figure 2: Preprocessed data of the examples shown in fig.1.

Some works proposed use of SVM (Malmasi and Zampieri, 2018), logistic regression (Davidson et al., 2017), random forest classifier (Burnap and Williams, 2015) and word embedding based method (Badjatiya et al., 2017) for classification of tweet contents. Liu et al. (2017) provided the use of unsupervised methods to cluster news topics from tweets.

The capability of dependency learning and semantic information extraction enables us to learn complex decision boundaries. The evolution around capturing the semantic relationships between words lead to the widely used Transformer (Vaswani et al., 2017) architecture.

Such models have outperformed conventional methods over natural language processing (NLP) tasks. They have been widely used for various real-world applications such as language modeling (Wang et al., 2019), sarcasm detection (Kumar Jena et al., 2020), summarization (Egonmwan and Chali, 2019), and other language tasks.

3 Dataset

The training and validation data consists of 7000 and 1000 samples, respectively. The test dataset consists of 12000 tweets, out of which 2000 tweets were selected by the organizers for final evaluation. The actual labels of the test dataset was not revealed by the shared task, hence the accuracy metric is only reported for the validation dataset in this paper. The number of samples in each class is presented in Table 1. The maximum, minimum, and average word count of the train and validation data is also shown in Table 1.

	Train	**Validation**
Number of samples in each class		
Informative	3303	472
Uninformative	3697	528
Word count - before preprocessing		
Maximum	76	62
Minimum	8	11
Average	35.87	37.052
Word count - after preprocessing		
Maximum	217	69
Minimum	7	10
Average	36.301	37.215

Table 1: Dataset statistics - number of samples in each classes, and word count before and after preprocessing data

Preprocessing data All the texts are converted to lower-case, and the emojis are replaced by their corresponding textual description. Further, contractions in the texts are fixed, and URLs and non-ascii characters are removed. It can be seen from Table 1 that the range of word count has increased after preprocessing for both the train and validation data. The word count increased due to emojis' conversion to text and decreased due to the removal of non-ascii characters. An example of a preprocessed tweet for both the informative and uninformative class is shown in Figure 2.

4 Methods

We have applied various conventional machine learning and transformer-based approaches.

4.1 Conventional approaches

We used traditional ways of word representation such as Bag-of-Words (BoW) and TF-IDF for detecting informative tweets using classifiers such as logistic regression, SVM, Naive Bayes, random forest classifier, and 2-layer MLP.

4.2 Transformer based approaches

Transformer is a way of improving the performance of NLP models. It is an encoder-decoder-type architecture that observes the whole of the input sequence at once. Unlike the recurrent sequential method, it uses an attention mechanism to detect long term dependencies. In this paper, we focus on experimenting with transformer-based architectures like BERT, DistilBERT, RoBERTa, and ALBERT.

Classifier	Bag-of-Words	TF-IDF Vectors
Logistic Regression	0.78318	0.78331
SVM	0.78054	0.78472
Naive Bayes	0.76371	0.74449
Random Forest	0.55489	0.56447
MLP	0.78695	**0.79912**

Table 2: F1 score of conventional approaches.

BERT Devlin et al. (2018) presents a bidirectional transformer-based language model, pre-trained on deep bidirectional representations from unlabeled text. It is jointly conditioned on both left and right context in all layers, and it is known for outperforming several state-of-the-art systems for various NLP tasks. We have used BERT-base-uncased pre-trained model to perform the informative tweet classification.

DistilBERT Sanh et al. (2019) proposes an approximate version of BERT using half the number of parameters. It improves the inference time while retaining 97% of the performance of BERT. The pre-trained model we used is DistilBERT-base-uncased to analyze the comparative classification with respect to the BERT model.

RoBERTa Liu et al. (2019) adopts the training mechanism used by BERT with a significantly longer training time over longer sequences. It differs from BERT as it uses a dynamic masking pattern compared to static in prior. We have used RoBERTa-base pre-trained model. It is trained with a significantly large dataset and outperforms BERT, DistilBERT, and other variants for various downstream tasks.

ALBERT Lan et al. (2019) introduces another light version of BERT, with low memory consumption and high training speed by which it outperforms the state-of-the-art models for various benchmark datasets. In our experiment, we used the pre-trained ALBERT-base-v1 model.

We have used all these pretrained models with the same parameters (discussed in section 5.2) except for RoBERTa and DistilBERT, which required small changes. In RoBERTa, before tokenizing the sentences, prefix space has to be set true, along with the addition of special tokens. In DistilBERT, the token type id's are not considered.

5 Implementation

5.1 Conventional approach

The BoW and TF-IDF vectors are obtained after the given training, and validation sets undergo preprocessing procedure. The liblinear solver is used in the logistic regression. The maximum depth of the decision tree in random forest classifier is set as 8. In the MLP classifier, lbfgs solver is used with alpha value set as 1e-5, the number of hidden layers is 2 with 5 and 2 neurons in the first and second layers, respectively. Other parameters concerning the conventional methods use default values.

5.2 Transformer-based approach

The sentences after undergoing the cleaning process are tokenized using the pre-trained model's tokenizer. Special tokens are added to detect the start and end of a sentence, and each token is mapped with an id. Next, the padding layer is added with value 0 and truncated to a maximum length of 100 to maintain equal lengths of the embeddings. Attention masks are used to detect padded tokens and actual words. Mask is set to 0 if the token id is 0, else it is set to 1.

The actual training set is further split into two parts (9:1 ratio), i.e., train and dev set to check which learning rate the model performs better. The input arguments are passed to evaluate our validation dataset. Finally, the F1 score is calculated between predicted and actual labels of the validation set.

Reproducibility We have considered batch size as 32, the learning rate of the optimizer as 2e-5, and its epsilon value is set as 1e-8. We trained our model for 4 epochs as determined by optimising on the dev set. To get the reproducible results, we set the seed value for all the Python packages. The torch seed, manual seed, and NumPy seed are set as 0. Further, while using CuDNN backend, we set deterministic as true and benchmark as false.

6 Results

The F1 scores on the validation dataset obtained for conventional methods and transformer-based methods are presented in Table 2 and Table 3 respectively.

It is observed that the TF-IDF vectorizer gives better results when compared to BoW in almost all the conventional approaches. Among all the conventional approaches, MLP gives the best result.

The transformer-based methods performed better compared to all the conventional approaches. BERT and RoBERTa showed competitive performance. However, RoBERTa has shown better results compared to other transformers based methods.

Classifier	F1 score
BERT	0.88634
ALBERT	0.87786
DistilBERT	0.88061
RoBERTa	**0.88991**

Table 3: F1 score of transformer-based approaches.

7 Conclusion

Classifying Twitter texts has been at the forefront of various NLP applications. Here, in this paper we have worked on one such task of classifying a tweet as informative or uninformative in the context of Covid-19. We have extensively compared the performance of various methods for this task. We applied conventional approaches and the latest state-of-the-art transformer-based methods. The results shows that the RoBERTa gives superior result on this task.

Since Twitter is primarily a microblogging media, short text classification using topic modeling (Zhang et al., 2013; Blei and Lafferty, 2009), and topic-enhanced embedding-based approach. (Li et al., 2016) can also be useful for tweet classification. In future work, we wish to apply topic modeling to produce enhanced word embeddings for this task.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188.

David M Blei and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

Pete Burnap and Matthew Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making: Machine classification of cyber hate speech. *Policy Internet*, 7.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

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*.

Elozino Egonmwan and Yllias Chali. 2019. Transformer-based model for single documents neural summarization. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 70–79, Hong Kong. Association for Computational Linguistics.

Amit Kumar Jena, Aman Sinha, and Rohit Agarwal. 2020. C-net: Contextual network for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 61–66, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. 2016. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2429–2432.

X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin, and J. Duprey. 2017. Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1483–1493.

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*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *CoRR*, abs/1803.05495.

Maria Martinez-Rojas, Maria del Carmen Pardo-Ferreira, and Juan Carlos Rubio-Romero. 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, 43:196–208.

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*.

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*.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chenguang Wang, Mu Li, and Alexander J. Smola. 2019. Language models with transformers. *CoRR*, abs/1904.09408.

Zhifei Zhang, Duoqian Miao, and Can Gao. 2013. Short text classification using latent dirichlet allocation. *Jisuanji Yingyong/ Journal of Computer Applications*, 33(6):1587–1590.