UPennHLP at WNUT-2020 Task 2 : Transformer models for classification of COVID19 posts on Twitter

Arjun Magge*
Perelman School of Medicine,
University of Pennsylvania,
Philadelphia, PA, USA
Arjun.Magge@pennmedicine.upenn.edu

Varad Pimpalkhute*
Electronics and Communication Engineering,
Indian Institute of Information Technology,
Nagpur, MH, India

Divya Rallapalli
Barrett Honors College,
Arizona State University,
Tempe, AZ, USA

David Siguenza
Great Valley High School,
Malvern, PA, USA

Graciela Gonzalez-Hernandez
DBEI, Perelman School of Medicine,
University of Pennsylvania,
Philadelphia, PA, USA
gragon@pennmedicine.upenn.edu

Abstract

Increasing usage of social media presents new non-traditional avenues for monitoring disease outbreaks, virus transmissions and disease progressions through user posts describing test results or disease symptoms. However, the discussions on the topic of infectious diseases that are informative in nature also span various topics such as news, politics and humor which makes the data mining challenging. We present a system to identify tweets about the COVID19 disease outbreak that are deemed to be informative on Twitter for use in downstream applications. The system scored a F1-score of 0.8941, Precision of 0.9028, Recall of 0.8856 and Accuracy of 0.9010. In the shared task organized as part of the 6th Workshop of Noisy User-generated Text (WNUT), the system was ranked 18th by F1-score and 13th by Accuracy.

1 Introduction

The COVID19 pandemic caused by the coronavirus (nCOV) has presented a unique challenge to the public health research community in the areas of tracking localized and community level transmissions for enforcing effective mobility restrictions and interventions to curb further virus spread. While traditional sources of infection numbers and fatalities include testing facilities and healthcare providers, many new non-traditional sources of information such as social media, wastewater analysis and mobility statistics that may serve as biomarkers for presence of infections in the community. In this work, we focus on using natural language processing (NLP) for mining social media posts for tweets that mention that can be used for public health monitoring purposes or dissemination of information. We accomplish this by introducing a classifier that can be used as a component of an information processing pipeline to detect informative tweets from posts that mention keywords related to discussions around coronavirus.

The system presented in this work was developed as part of the W-NUT 2020 shared task 2 (Nguyen et al., 2020b) where the objective of the task was to classify a given post as informative or uninformative. The rest of the document is structured as follows: we briefly discuss previous related work on COVID19 surveillance on Twitter in the Background section. We describe the annotated dataset and system implementation in the Materials and Methods section followed by preliminary and final evaluation results in the Results section. Finally, we discuss error analysis, limitations and future directions in the Discussion section.

2 Background

Researchers working on noisy user texts such as posts on social media mining have proposed various methods and systems for monitoring infectious disease transmissions and natural disasters

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 378–382
Online, Nov 19, 2020.

Corpus	Informative	Uninformative
Training Set	3303	3697
Validation Set	472	528
Test Set	944	1056

Table 1: Dataset description for identifying informative tweets on Twitter

such as hurricanes (Paul and Dredze, 2017). Most work on COVID19 monitoring have focused on presenting keywords for data collection related to COVID19 (Chen et al., 2020; Rashed et al., 2020; Wei et al., 2020; Santosh et al., 2020), datasets for classification of tweets into categories for downstream applications(Klein et al., 2020; Golder et al.; Delizo et al., 2020; Liu et al., 2020; Karisani and Karisani, 2020; Müller et al., 2020; Jelodar et al., 2020; Lwowski and Najafirad, 2020; Mackey et al., 2020), and in some cases advanced tasks such as event detection (Zong et al., 2020), detection of symptoms experienced by users who tested positive for the disease (Al-Garadi et al., 2020) and review articles on the topic (Arafat, 2020; Moore et al., 2020).

The classification categories themselves have varied from sentiments expressed in posts (Delizo et al., 2020) and relatedness to the disease (Liu et al., 2020; Karisani and Karisani, 2020) to misinformation detection (Hossain et al., 2020) and personal reports of exposures or test results (Klein et al., 2020). Each dataset contains tweets and annotations that can be processed by information processing pipelines using NLP and machine learning based classification techniques for possible applications in public health using epidemiological analysis.

3 Materials and Methods

The dataset annotated for the task consists of 10,000 tweets that mentioned terms related to COVID19. This included a total of 4719 tweets labeled as *Informative* and 5281 tweets labeled as *Uninformative*. Each tweet was annotated by 3 independent annotators with an inter-annotator agreement score of Fleiss' Kappa of 0.818. The dataset was split into training set (70%), validation set (10%) and test set (20%) for the purposes of development and evaluation of the classification model. We tabulate further details of the splits in Table 1.

Architecture	Prec	Recall	F1	Acc
LR	0.85	0.80	0.82	0.83
Feedforward	0.82	0.80	0.80	0.79
CNN	0.86	0.88	0.84	0.85
BERT	0.83	**0.96**	0.89	0.89
GPT	0.84	0.89	0.86	0.87
XLNET	0.82	0.92	0.90	0.90
RoBERTa	0.86	0.93	0.89	0.89
DistilBERT	0.83	0.93	0.88	0.88
BERTweet	0.88	0.90	0.89	0.90
BERT-Epi	**0.91**	0.93	**0.92**	**0.92**

Table 2: Performance of the experimentation systems on the validation set. Precision, recall and F1-score was calculated for the *Informative* class. We used the above scores to determine the final model used for making the official submission.

3.1 Pre-processing

We pre-processed each tweet to normalize usernames and urls into reserved keywords. Further, we de-emojized the tweets using the emoji package to add descriptive lexical features that convey emotions associated with the tweet. Finally, we expanded contractions for normalizing the text for detecting negations.

3.2 Classification models

We experimented with various machine learning models such as logistic regression with bag-of-word features, convolutional neural networks (CNN) with filter sizes upto 5, fully connected (feed-forward) network, and transformer models using the scikit-learn [1], TensorFlow (Abadi et al., 2016), ktrain (Maiya, 2020) and Flair (Akbik et al., 2018) frameworks.

3.2.1 Non-transformer models

In order to setup baselines, we used the scikit-learn framework to convert the tweet into count vectors based on the individual tokens in the tokenized sentences. For fast inference, we setup the logistic regression (LR) baseline using the count vectors as features. We did not perform elaborate experiments such as employing various word embeddings, n-gram features and feature engineering in favor of building models with better classification performance. We used the same feature set for training a fully connected Feedforward classifier using Tensorflow keras API with 20 neurons and a sigmoid

[1] https://scikit-learn.org/

layer as an output layer. Lastly we built a CNN text classification model with filter sizes upto five using the same input features discussed above.

3.2.2 Transformer models

We also experimented with various transformer language models such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), GPT (Radford et al.), XLNET (Yang et al., 2019) and RoBERTa (Liu et al., 2019) among others that were specifically trained on COVID19 related datasets such as BERT-Epi(Müller et al., 2020), and BERTweet (Nguyen et al., 2020a). We used the Flair framework for training the classifier model where all layers of the model were fine-tuned during the training. The performance was measured across the standard classification metrics of precision, recall and F1-score with the final determining metric identified as the F1-score for the *Informative* class.

We trained each of these models with Adam optimizer and a softmax layer with weighted losses such that *Informative* class loss was weighted 2 times relative to the loss of the *Uninformative* class. The full comparison of the performance of these models has been shown in Table 2. We found that various ensembles of best combinations did not result in better models and hence we chose the final transformer model that was an uncased BERT model trained on COVID19 related tweets (Müller et al., 2020) trained using the Flair framework (Akbik et al., 2018). After hyperparameter tuning the model, we determined the best parameters to be a learning rate of 0.00003 and loss weights of 2 in favor of *Informative* class and the usage of Adam optimizer. For the final model we trained the model on all available tweets combining training and validation sets for 10 epochs. All training experiments including development of the final submission model was performed on Google Colaboratory [2] .

4 Results

The final detailed results of the task on the validation and test sets are shown in Table 3. Observing the results of the validation and test set, we find that the performance of the final model deteriorated on the test set. The drop in recall on the test set was significant even though the splits between the classes remained the same. The final model was ranked 18th by the F1-score and 13th by accuracy. The final standings relative the proposed system is shown in Figure 1. We find that the performance of most of the top models (ranks 1-20) including the proposed system were determined to have F1-scores within a narrow margin of 0.89 and 0.91 which shows that the shared task was very competitive.

[2]https://colab.research.google.com/

Corpus	Prec	Recall	F1	Acc
Validation	0.894	0.936	0.914	0.918
Test	0.902	0.885	0.894	0.901

Table 3: Performance of the system on the validation and test set for identifying informative tweets on Twitter.

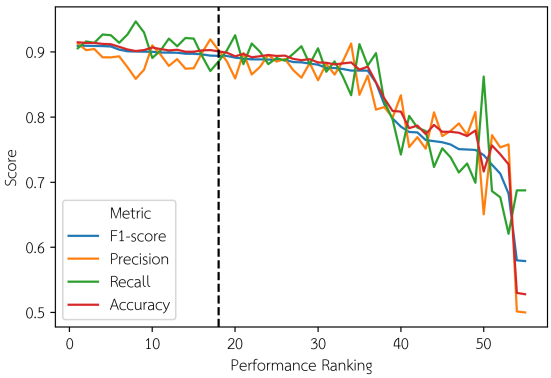

Figure 1: Performance of the systems submitted as part of the shared task across the metrics of F1-score, Precision, Recall and Accuracy compared to the system presented in our work indicated by the dotted line.

5 Discussion

On closer analysis of initial errors on the Validation set we found that many tweets were difficult to determine as being in either classes. On performing an 8-fold cross validation, we found that fold 1 which the default split of training and validation set had the lowest F1-scores at 0.92 whereas folds 2-8 had F1-scores in the ranges of 0.94-0.96. On closer analysis of fold 1, we found that many tweets may have been annotated incorrectly in the validation set which may have introduced errors in the final model and hence may require further analyses of the annotations for development of better models.

Overall, transformer models performed significantly better than non-transformer models that we trained on. However, we note that the non-transformer models could be improved by using word-embeddings and attention features in layers among Feedforward and CNN architectures.

The current system proposes a simple classification method which may be useful in removing tweets that are deemed *Uninformative* for use in downstream epidemiological analyses. Further research is required to assess the utility of the tweets obtained from such collections of *Informative* tweets for disease tracking and analyses of symptoms.

5.1 Limitations

Some of the common drawbacks for performing demographic analyses include the problem of selection bias in Twitter users where most of the users tend to live in cities and possess smartphones which may not be representative of the overall population of countries or individual administrative regions. Another drawback noticed among systems trained on social media data is that the model performance seems to decline over time as new terms are introduced into the vocabulary and newer topics are mentioned in conversations around the disease. Although, tweets publicly published on Twitter are available for viewing by users and non-users of Twitter, development of automated methods to determine posts that may determine infections/diagnoses which may be personal in nature may raise valid privacy concerns due to possible adverse social impacts on such individuals. Such information-sharing policies may warrant constant review with changing times and national policies. In this paper, we do not include direct text content of tweets during error analysis for aforementioned reasons.

6 Conclusion

Social media mining offers a non-traditional avenue to extract epidemiological and population level statistics from discussions around a topic. Such noisy information domains presents an NLP challenge for extracting meaningful information for use in downstream applications. In this work, we present a system to identify *Informative* tweets on English language Twitter posts on the topic of COVID19 pandemic for use in downstream tasks such as public health monitoring and epidemiological studies. We use a classification approach to identifying such tweets and the final system scored a F1-score of 0.8941, Precision of 0.9028, Recall of 0.8856 and Accuracy of 0.9010. In the shared task organized as part of the 6th Workshop of Noisy User-generated Text (WNUT), the system was ranked 18th by F1-score and 13th by Accuracy.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2020. A text classification approach for the automatic detection of twitter posts containing self-reported covid-19 symptoms.

Mahmoud Arafat. 2020. A review of models for hydrating large-scale twitter data of covid-19-related tweets for transportation research.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

John Pierre D Delizo, Mideth B Abisado, and Ma Ian P De Los Trinos. 2020. Philippine twitter sentiments during covid-19 pandemic using multinomial naïve-bayes. *International Journal*, 9(1.3).

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Su Golder, Ari Z Klein, Arjun Magge, Karen O'Connor, Haitao Cai, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. Extending a chronological and geographical analysis of personal reports of covid-19 on twitter to england, uk. *medRxiv*.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sameer Singh, and Sean Young. 2020. Detecting covid-19 misinformation on social media.

Hamed Jelodar, Yongli Wang, Rita Orji, and Hucheng Huang. 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *arXiv preprint arXiv:2004.11695*.

Negin Karisani and Payam Karisani. 2020. Mining coronavirus (covid-19) posts in social media. *ArXiv*.

Ari Klein, Arjun Magge, Karen O'Connor, Haitao Cai, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. A chronological and geographical analysis of personal reports of covid-19 on twitter. *medRxiv*.

Junhua Liu, Trisha Singhal, Lucienne Blessing, Kristin L. Wood, and Kwan Hui Lim. 2020. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. *ArXiv*, abs/2005.06627.

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*.

Brandon Lwowski and Peyman Najafirad. 2020. Covid-19 surveillance through twitter using self-supervised learning and few shot learning.

Tim Mackey, Vidya Purushothaman, Jiawei Li, Neal Shah, Matthew Nali, Cortni Bardier, Bryan Liang, Mingxiang Cai, and Raphael Cuomo. 2020. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with covid-19 on twitter: Retrospective big data infoveillance study. *JMIR Public Health and Surveillance*, 6(2):e19509.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv*, arXiv:2004.10703 [cs.LG].

Jason H Moore, Ian Barnett, Mary Regina Boland, Yong Chen, George Demiris, Graciela Gonzalez-Hernandez, Daniel S Herman, Blanca E Himes, Rebecca A Hubbard, Dokyoon Kim, et al. 2020. Ideas for how informaticians can get involved with covid-19 research.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020a. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

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

Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Salma Rashed, Johan Frid, and Sonja Aits. 2020. English dictionaries, gold and silver standard corpora for biomedical natural language processing related to sars-cov-2 and covid-19.

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*.

Roshan Santosh, Sharath Chandra Guntuku, H Schwartz, Lyle Ungar, et al. 2020. Detecting symptoms using context-based twitter embeddings during covid-19.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What are people asking about covid-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. Extracting covid-19 events from twitter. *arXiv preprint arXiv:2006.02567*.