#GCDH at WNUT-2020 Task 2: BERT-Based Models for the Detection of Informativeness in English COVID-19 Related Tweets

Hanna Varachkina
University of Göttingen
GCDH*
Papendiek 16
37073 Göttingen
Germany
hanna.varachkina@stud.uni-goettingen.de

Stefan Ziehe
University of Göttingen
Institute of Computer Science
Goldschmidtstraße 7
37077 Göttingen
Germany
stefan.ziehe@cs.uni-goettingen.de

Tillmann Dönicke
University of Göttingen
GCDH*
Papendiek 16
37073 Göttingen
Germany
tillmann.doenicke@uni-goettingen.de

Franziska Pannach
University of Göttingen
GCDH*
Papendiek 16
37073 Göttingen
Germany
franziska.pannach@stud.uni-goettingen.de

Abstract

In this system paper, we present a transformer-based approach to the detection of informativeness in English tweets on the topic of the current COVID-19 pandemic. Our models distinguish informative tweets, i.e. tweets containing statistics on recovery, suspected and confirmed cases and COVID-19 related deaths, from uninformative tweets. We present two transformer-based approaches as well as a Naive Bayes classifier and a support vector machine as baseline systems. The transformer models outperform the baselines by more than 0.1 in F1-score, with F1-scores of 0.9091 and 0.9036. Our models were submitted to the shared task *Identification of informative COVID-19 English tweets (WNUT-2020 Task 2).*

1 Introduction

As of the end of August 2020, the outbreak of the COVID-19 pandemic has lead to 24.5 million cases world-wide and affected individuals, communities and nations in all parts of the world[1]. In order to receive reliable data on cases in their area, users rely on information systems, such as the Coronavirus Resource Center introduced by the Johns Hopkins University.[2] However, those systems mainly utilize statistics provided by official health institutes, such as the World Health Organisation or the Robert Koch Institute. In order to confirm the numbers of COVID-19 related confirmed cases, recoveries or deaths, those official information channels release data once a day.

Alternative sources, e.g. social media outlets such as Twitter, publish a vast amount of data in real-time. Therefore, the WNUT-2020 Shared Task 2 (Nguyen et al., 2020) called for systems that can distinguish informative Twitter messages, so-called tweets, from uninformative data. A sophisticated informativeness detection system can be used to update COVID-19 information systems for users world-wide on a real-time basis.

In this paper, we present different machine-learning approaches on the identification of informative tweets. We first evaluate Naive Bayes and support-vector-machine approaches with hand-crafted features. Subsequently, we employ transformer-based models, such as COVID-Twitter-BERT (Müller et al., 2020), which provide the best results.

This paper is structured as follows: Section 2 explains the task and the data provided by the hosts of the shared task, Section 3 explains different approaches on the machine-learning systems, Sections 4 and 5 present and discuss results. We finish this paper with a conclusion in Section 6 and an outview on future work in Section 7.

Our source code is available at https://gitlab.gwdg.de/tillmann.doenicke/wnut-2020.

2 Data

The organizers of the shared task provided raw data of a tweet ID, tweet text and the label "informative" or "uninformative". Participants were free to make use of additional data (Nguyen et al., 2020). We used the suggested split of the dataset with 7,000 tweets for training and 1,000 for validation.

3 Models

3.1 Baselines

For our baseline models, we tokenized the tweets with NLTK's tweetTokenizer[3], lowercased all tokens except all-capitals tokens and optionally

*Göttingen Centre for Digital Humanities

[1] https://coronavirus.jhu.edu/map.html

[2] https://coronavirus.jhu.edu/

[3] https://www.nltk.org/api/nltk.tokenize.html

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 462–465
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics

masked all numbers except “19” (as in “COVID-19”). (Masking numbers means replacing them with a special token “#NUMBER#”.) We then extracted token n-grams with $n \in \{1, \dots, 5\}$ to use them as features.

Our first baseline model is a Naive Bayes classifier (NBC) with binary features (n-gram in tweet or not), implemented with NLTK’s NaiveBayesClassifier[4].

The second baseline model is a support vector machine (SVM), implemented with scikit-learn’s SVM module[5]. We tested binary features, count features (frequency of n-gram in tweet) and tf·idf features. To reduce computation time, we only considered n-grams which occurred at least 10 times in the training corpus.

We also tested a wide range of additional features for the SVM: a tweet’s sentiment analysis scores, its Gunning fog index, the number of medical terms (we extracted the most informative terms from the CORD-19 dataset[6], a collection of scholarly articles about COVID-19), most of the content features[7] from Wang et al. (2015), and Boolean features which capture whether a tweet ends with an URL (to capture posts with a link to a source, e.g. a newspaper article), starts with a USER (to capture retweets), or contains an opinion expression from a hand-crafted list such as “I think” or “my point of view”.

Suspecting that informativeness is expressed through statements with neutral sentiment, we extracted sentiment using the NLTK sentiment analysis module VADER[8]. However, we assume that the majority of the raw data provided in the task was already pre-filtered for neutral sentiment, as illustrated in Figure 1: The smallest score for neutral sentiment observed in the raw training data set was 0.335 (1 tweet), with 998 tweets that showed a perfect neutral sentiment score of 1.

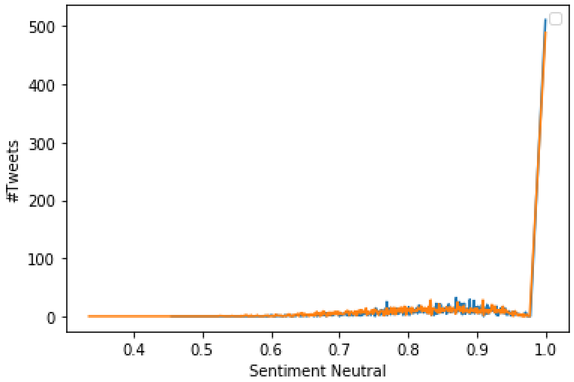

Figure 1: Distribution of neutral sentiment in informative (orange) and uninformative (blue) tweets, 1 being perfectly neutral

3.2 BERT-Based Models

Our other models rely on BERT (Devlin et al., 2018), a language model based on the transformer architecture that can be adapted to text classification tasks. For these models the data was preprocessed using the code provided by Müller et al. (2020). Retweet tags were removed, user names and URLs were replaced by a corresponding text token (“twitteruser” or “httpurl”) and all unicode emoticons were substituted by textual ASCII representations. These preprocessing steps converted 64 tweets into empty strings, which were subsequently removed from the training set. We further moved the last 24 tweets from the training set (now 6,912 tweets) to the validation set (now 1,024 tweets). This allows for using a constant batch size of 32 during training.

The first model is based on the uncased BERT-large model. We fine-tuned it for two epochs on the training data for the given classification task. The second one is based on COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020), a version of the uncased BERT-large model that was fine-tuned on 22.5M COVID-19 related tweets (0.6B words). CT-BERT was our choice because it is also based on BERT-large and is specifically fine-tuned for tweets related to COVID-19. We further fine-tuned it for three epochs on the training set. All the training was done on Tensor Processing Units (TPUs) provided in Google Colaboratory[9], which allowed us to use a sequence length of 96, a training batch size of 32 and an evaluation batch size of 1024. We used the Adam optimizer (Kingma and Ba, 2015) with a base learning rate of 2e-5. Additionally, we

[4] http://www.nltk.org/api/nltk.classify.html#module-nltk.classify.naivebayes

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[6] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[7] Number of words, number of characters, number of capitalization words, number of capitalization words per word, maximum word length, mean word length, number of exclamation marks, number of question marks, number of URL links, number of URL links per word, number of hashtags, number of hashtags per word, number of mentions, number of mentions per word.

[8] https://www.nltk.org/_modules/nltk/sentiment/vader.html

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

used smaller learning rates for earlier BERT layers by decaying them with a decay factor of $\xi = 0.95$ as described in Sun et al. (2019).

4 Results

Table 1 shows the results for all models. The NBC achieves its highest F1-score of 0.8012 when using only uni- and bigrams as features and not masking numbers. The SVM achieves its highest F1-score of 0.8009 when using uni- and bigrams as binary features and masking numbers. None of the additional features did help improving the performance of the SVM.

BERT-based models perform better than baseline models. The best result was achieved using the CT-BERT model trained on the domain-specific Twitter data on the topic of COVID-19 ("CT-BERT-1" in Table 1). It reached an F1-score of 0.9231 on the validation set. In comparison, the BERT-large model only achieved an F1-score of 0.8928.

Model	F1	Prec.	Rec.	Acc.
NBC	0.8012	0.7721	0.8326	0.8050
SVM	0.8009	0.8380	0.7669	0.8200
BERT	0.8928	0.8965	0.8891	0.8984
CT-BERT-1	0.9231	0.8988	0.9487	0.9248

Table 1: F1-score, Precision, Recall and Accuracy on the validation set

We submitted predictions made by CT-BERT to the shared task, both after training only on the training set (CT-BERT-1) for three epochs and after additionally training on the validation set (CT-BERT-2) for two epochs. We achieved similar results in both cases. A larger number of training examples (7,936 compared to 6,912) could not improve the model's performance on the final test set (Table 2), although it led to a higher precision (Table 3).

Model	Training data	F1
CT-BERT-1	6,912 tweets	0.9091
CT-BERT-2	7,936 tweets	0.9036

Table 2: Model overview and F1-score on the final evaluation set

Note that the final test set of about 2,000 tweets was hidden in a larger test set of about 12,000 tweets for the system evaluation (Nguyen et al., 2020). Numbers reported in Tables 2 and 3 are provided by the shared task organizers.

Model	Prec.	Rec.	Acc.
CT-BERT-1	0.8919	0.9269	0.9125
CT-BERT-2	0.9036	0.9036	0.9090

Table 3: Precision, Recall and Accuracy on the final evaluation set

5 Discussion

Due to the fact that the transformer-based models were pre-trained on much larger, external datasets (Devlin et al., 2018; Müller et al., 2020), it is unsurprising that they outperform our baseline models, which were solely trained on the shared task's dataset.

Our transformer-based models predict informative tweets slightly better than uninformative tweets. On the validation set, the BERT model delivers 115 falsely classified tweets, 28 of which were wrongly classified as uninformative and 87 which were classified as informative. The CT-BERT-1 model falsely predicted 25 tweets as uninformative and 52 as informative. Therefore, the main difference lies in the prediction of uninformative tweets, for which CT-BERT-1 yields better results. This leads to a better recall of the CT-BERT-1 model compared to the BERT model.

Upon manual inspections, the falsely classified tweets fall into different categories, such as tweets that show statistics but also a personal opinion, personal stories and events, or political statements and protective measures against COVID (e.g. "amazon and facebook ask seattle employees to work from home after coronavirus cases httpurl" annotated as informative). We cannot determine a pattern for the cases in which CT-BERT-1 yields better predictions than BERT.

Without the guidelines on how tweets were annotated for the dataset, it is hard to assess why certain tweets were initially labelled as informative or uninformative in the data provided. Therefore, we were not able to identify the reason why certain predictions were incorrect.

6 Conclusion

We show in our paper how a transformer-based approach can yield good results with relatively little fine-tuning.

Our best-performing model, the CT-BERT-1 model, predicted informativeness with an F1-score of 0.9091 in the final W-NUT Task 2 evaluation.

Detecting informativeness of a small text unit, such as a tweet, is one of the many steps that are required in order to build trustworthy real-time news applications.

7 Future Work

The task of identifying informativeness in small text units, such as Twitter messages, is an important step towards building real-time news applications. In order to build systems that can be trusted, the next step needs to address the validation of claims. Verifying how factual a statement is, requires additional knowledge bases and ultimately, a common, indisputable understanding of what "the truth" is.

Furthermore, such a system needs to consider that a piece of information might be correct only at a certain point in time. A Twitter message might be both informative and factual, but provide information that is outdated, such as yesterday's number of new COVID-19 cases, which ultimately has little value in terms of real-time news feeds.

Solving the binary classification task of distinguishing informative tweets from uninformative tweets does not yet determine whether a tweet contains relevant information.

Lastly, the common goal of researchers in the field of informativeness and factuality detection should be to create resources that are general enough to be used without a large amount of event related data or human annotations. Only then we are able to provide systems that can be employed immediately for new arising crisis situations, such as natural disasters.

References

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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, 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*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics*, pages 194–206. Springer International Publishing.

Bo Wang, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2015. Making the most of tweet-inherent features for social spam detection on Twitter. *arXiv preprint arXiv:1503.07405*.