CXP949 at WNUT-2020 Task 2: Extracting Informative COVID-19 Tweets - RoBERTa Ensembles and The Continued Relevance of Handcrafted Features

Calum Perrio
School of Computer Science
University of Birmingham
United Kingdom
cperrio2015@gmail.com

Harish Tayyar Madabushi
School of Computer Science
University of Birmingham
United Kingdom
harish@harishtayyarmadabushi.com

Abstract

This paper presents our submission to Task 2 of the Workshop on Noisy User-generated Text. We explore improving the performance of a pre-trained transformer-based language model fine-tuned for text classification through an ensemble implementation that makes use of corpus level information and a handcrafted feature. We test the effectiveness of including the aforementioned features in accommodating the challenges of a noisy data set centred on a specific subject outside the remit of the pre-training data. We show that inclusion of additional features can improve classification results and achieve a score within 2 points of the top performing team.

1 Introduction

Identification of informative tweets in relation of coronavirus presents a text classification problem. Pre-trained bidirectional transformer-based models such as such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have proven to be extremely successful on text classification tasks; the process of pre-training on a large corpora enables the generation of effective contextual embeddings during fine-tuning, which can be leveraged on the classification task through the addition of an output layer (Devlin et al., 2018). The progression of the state-of-the-art that these models have facilitated is clearly demonstrable on performance against the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019); a collection of Natural Language Understanding tasks with an associated online platform for evaluation and analysis.

The corpora utilised during pre-training of transformer-based models typically consists of documents written in formal English. Moreover, this data is highly unlikely to contain references to the coronavirus pandemic which has only just occurred. Therefore, the noise inherent in social media data and subject specificity to coronavirus in the current data set (Nguyen et al., 2020), present challenges to conducting text classification with a pre-trained transformer-based model alone. To this regard, supplementary information may be useful to improving performance; corpus level information may potentially capture notions of relevance to words which fall outside the pre-trained vocabulary, and handcrafted features may show additional distinctions between classes. In this paper, we describe the development of a system with includes such features through the use of an ensemble, and which in the final submission to the evaluation stage achieved an F_1 Score of 0.8910.

The rest of this paper is organised as follows, firstly a discussion of related work is presented at Section 2. This is followed by a description of the methodological approach at Section 3. We present our results and analysis at Section 4 and present our conclusions in Section 5.

2 Related Work

As the 6th Workshop of Noisy User-generated Text presents the first time Task 2 has been made available, we consider research utilising pre-trained bidirectional transformer-based models from similar tasks to better enable informed decision making.

2.1 Related Tasks

SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (Zampieri et al., 2020), presented at sub-task A a similar binary classification problem to the present task. Many of the highest performing teams on this sub-task made use of "contextualised BERT-style Transformers" (Zampieri et al., 2020). Sub-task A specifically comprised of identifying whether a tweet presented content that contained inappropriate language, threats or insults, or was neither offensive

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 352–358

nor profane (Zampieri et al., 2020). Additionally, the sub-task was split based on the language of the tweet text, and to this regard we only consider relevant systems on the English data set, which present the highest degree of correlation to Task 2.

Wiedemann et al. (2020) achieved first place on SemEval-2020 Task 12 sub-task A utilising an ensemble of RoBERTa (Liu et al., 2019). Additionally, the authors leveraged the Masked Language Model pre-training objective of RoBERTa and further pre-trained on an *in-domain* data set. Domain specific pre-training is observed to be a method that is established as improving later results of supervised task-specific fine-tuning. Further pre-training is equally explored by Sotudeh et al. (2020) in their submission on SemEval-2020 Task 12 sub-task A that utilised BERT and achieved 4thplace.

Lim and Madabushi (2020) presented an ensemble model of BERT and TF-IDF. They hypothesise that corpus level count information captured by TF-IDF can boost the performance of BERT, and the authors achieve a result within 2 points of the top scoring team. Moreover, they note this performance was achieved using only 10% of the available training data due to physical constraints. We implement a similar method of incorporating TF-IDF features described at Section 3, but differentiate from this work through our ability to incorporate the entire training data.

2.2 User-generated Content

A key aspect of the present task to consider is the nature of Twitter data, or more broadly User-generated Content. This is specifically relevant to our implementation using transformer based models, with Kumar et al. (2020) showing the reduction in the performance of BERT with the introduction of noise in the form of spelling mistakes and typos. Similar inaccuracies can be expected in the data set for the present task given the informal nature of Twitter, the distinctive character limit, and modification of words by user to indicate emotion.

In relation to overcoming the challenges of Twitter data specifically, Ying et al. (2019) leveraged a token pattern detector to obtain domain-specific features. This comprised a convolution network trained on annotated features derived from pre-processing for domain-specific features. These representations were concatenated with the [CLS] token embedding from BERT, with the resulting model producing a statistically significant improvement on multi-label emotion classification over pure BERT (Ying et al., 2019).

2.3 Leveraging Metadata

Social media's widespread use generates a vast amount of data, not just in text but additionally in metadata. In the body of work surrounding the classification of rumors on Twitter , metadata has been successfully leveraged to develop handcrafted features (Li et al., 2019b). In the classification model produced by Li et al. (2019a), metadata information specific to Twitter, such as whether the account id is verified, if the profile includes a location or a description is concatenated together to form a input of "User Information" which is passed to a classifier along with textual and other features (Li et al., 2019a). This methodology of handcrafting features which provide supplementary information to a classifier is particularly related to the models in this work where we leverage a handcrafted feature derived from the text (Section 3).

3 Methodology

The data for this work was provided by the Workshop on Noisy User-generated Text (Xu et al., 2019) and consisted of 10,000 English tweets in relation to Covid-19. A 70/10/20 rate had been used to split the 10,000 tweets into training, validation and test sets. Each tweet had been labelled either as 'uninformative' or 'informative' by three independent annotators with an inter-annotator agreement score of Fleiss' Kappa at 0.818 (Xu et al., 2019). At the onset of this work, only the training and validation data had been provided. The test set was retained as a holdout set for the Workshop to evaluate the performance of models, discussed at Section 5.

In an attempt to build upon the success of leveraging tweet metadata in a classification model (see Section 2.3) an initial exploration of the distribution of metadata and text features between the "uninformative" and "informative" classes in the training dataset was conducted. This processes highlighted little differentiation between the classes, suggesting the distinction was more nuanced and conveyed within the circumstantial and contextual meanings. However, it was found that the higher the probability any given character in tweet was a numeric character increased the probability of it being "informative". We present a visual representation of this at Figure 1.

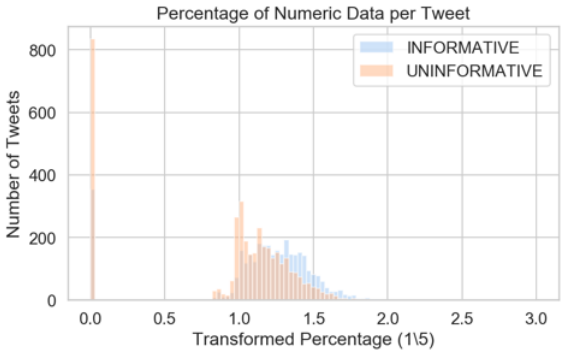

Figure 1: Probability of a character being numeric (power transformed $x^{\frac{1}{5}}$).

By virtue of the competitive element, this task presented an incentive to find the best performing model. In order to have a basis for comparison two initial models were trained on the raw training data and evaluated against the validation data: **(1)** a Linear Support Vector Machine (SVM) using TF-IDF features, and **(2)** fine-tuned $\text{BERT}_{\text{BASE}}$. As is the case with all the deep networks tested in this study, we test a range of hyperparameter values (details of which are available in the program code and model details released as part of this publication [1]), and test each combination against five different random seeds. The effect of random seed on BERT is emphasised in Dodge et al. (2020) where varying only the random seed was shown to produce substantial improvements over previously published results. We present the highest performing results of these models at Table 1.

Model	F_1 Score
SVM	0.8155
BERT	0.9051
RoBERTa (unprocessed)	0.9101
RoBERTa (processed)	0.9131

Table 1: Baseline results achieved by BERT and SVM, and results from RoBERTa. All tested aganist the validation data.

The SVM model was chosen due to the established high accuracy this model can achieve on the task of text classification ((Kadhim, 2019); (Shah and Patel, 2016)).

We observe that a fine-tuned BERT model achieved a good performance against the validation data. The next logical progression was to experiment with a fine-tuned RoBERTa implementation, given that the robustly optimised pre-training this model employs improved performance over BERT (Liu et al., 2019). We fine-tuned a $\text{RoBERTa}_{\text{BASE}}$ model, and the results of this are presented against the BERT and SVM baselines in Table 1. We present two versions of the RoBERTa model: processed and unprocessed. The processed implementation draws inspiration from the pre-processing adopted by Nikolov and Radivchev (2019), in which we also segmented camel-cased hashtags into distinct tokens e.g. the token "#HashTag" would become "#Hash" and "Tag". Additionally, the two forms primarily used to refer to coronavirus: "covid-19" and "coronavirus" were parsed for in an extensive number of variations and standardised to "coronavirus" to ensure the token representation for this key term was consistent across inputs.

Model	False Negative	False Positive
BERT	62.16%	41.81%
RoBERTa	56.6%	43.27%

Table 2: Percentage of miss-classifications shared with the SVM. Visual representations are included in the Appendix.

An error analysis was conducted after the development and testing of RoBERTa which explored the intersection of the false negative and false positive classifications from the SVM, BERT and the processed RoBERTa implementations. We present the most interesting finding from this in Table 2: despite higher overall miss-classification, the SVM trained on TF-IDF features was able to correctly classify a significant proportion of miss-classifications from the fine-tuned BERT and RoBERTa models. This suggests, as in Lim and Madabushi (2020), that incorporation of corpus level information could improve performance.

Additionally, the error analysis noted the BPE tokenization of "coronavirus" was out-of-vocabulary, splitting into the tokens "Ġcoron, av, irus". To this regard, post submission we have tested a promising implementation that instead replaces all forms referring to "coronavirus" to "coronavirus disease". With the intention of leveraging through self-attention the context of these terms as a disease. This is explored further in the future work.

3.1 Exploration of Improving Baseline Performances

The error analysis and initial feature exploration motivated the experimentation of improving the performance of a pre-trained language model in the

[1] https://github.com/CalumPerrio/WNUT-2020

following three ensemble models. In all instances the RoBERTa processed implementation was used as the base pre-trained language model due to it achieving the highest F_1 score on the validation data in our initial experiments.

Three ensemble models were experimented, presented broadly below and explored in details at Section 3.2, these were: **a)** RoBERTa together with a percentage metric of the probability of a character in the text being numeric, **b)** RoBERTa with TF-IDF features, and **c)** RoBERTa with both TF-IDF features and the percentage metric.

3.2 Model Architecture

For all three ensemble models, we concatenate the additional features to the final hidden layer for the $< /s >$ token, which we use as an aggregate representation of the sequence. This vector then forms the input to the fully connected output layer.

With regards to the TF-IDF features. A document-term matrix was constructed from a processed version of the training data, and a TF-IDF feature vector fit for the text for the equivalent tokenized input to the model. Processing involved removing punctuation, stopwords and emojis, and stemming. During testing, two combinations of maximum features in the document-term matrix were tested: 6000 and 9000.

The probability metric was produced by first removing emojis from the text to prevent the influence of numbers in unicode. Then returning the probability of any character in the tweet being a digit, transformed as a (1,1) tensor.

4 Results and Analysis

We present the results of the three ensemble models in comparison to the processed RoBERTa model from Table 1 in Table 3.

Model	F_1 Score
RoBERTa (baseline)	0.9131
RoBERTa+PROB	0.9141
RoBERTa+TFIDF	0.9111
RoBERTa+PROB+TFIDF	0.9151

Table 3: F1 score results of the three ensemble models presented against the baseline performance of RoBERTa.

We observe that against the baseline RoBERTa model, inclusion of the hand-crated feature improved performance. However, the increase is small, potentially reflecting that the number of tweets in the data set for which this features is a useful indicator is equally small. We also observe that the inclusion of TF-IDF features on the validation data alone did not improve performance, however the inclusion of both the average metric and TF-IDF produced an improved result over the baseline and RoBERTa+PROB. The former result does not support our hypothesis of improving performance with the inclusion of TF-IDF features and is in contradiction to the latter result.

Between the RoBERTa+TFIDF model and RoBERTa+TFIDF+PROB models, the highest performances were found on models featuring 9000 and 6000 TF-IDF features respectively. It is notable that the total dimension of the stemmed document-term matrix was over 15,000 features. As such, due to memory constraints it was not possible to use a vector of this entire dimension. We submit an interesting experiment would be to parse for domain-specific expressions as in (Ying et al., 2019) and standardise these to dictionary representation to reduce the number of TF-IDF features. Additionally, standardisation of these expressions across the corpus could potentially enable greater accuracy in calculating the Inverse Document Frequency, and by virtue of this better capture the amount of information a term provides.

In our submission to the evaluation stage we submit the RoBERTa+PROB+TFIDF ensemble model which achieved the highest performance of all models tested against the validation data (see Table 3). Our results against the holdout test set are presented at Table 4 [2]. Our ensemble model achieves an F_1 score of 0.8910, within two points of the highest performing team.

Rank	Team	F1 Score
1	NutCracker	0.9096
2	NLP_North	0.9096
...		
21	cxp949 (this work)	0.8910
...		
55	TMU-COVID19	0.5000

Table 4: F1 score of Final Model against the holdout test set.

4.1 Error Analysis

We present at Figure 2 the confusion matrices of the baseline RoBERTa model and RoBERTa+PROB+TFIDF ensemble model. We

[2]Independently evaluated by organisers.

observe that inclusion of the handcrafted feature and TF-IDF features provided a net gain on true-positive "uninformative" classifications, and reduced the number of false-positive miss-classifications when compared with the baseline.

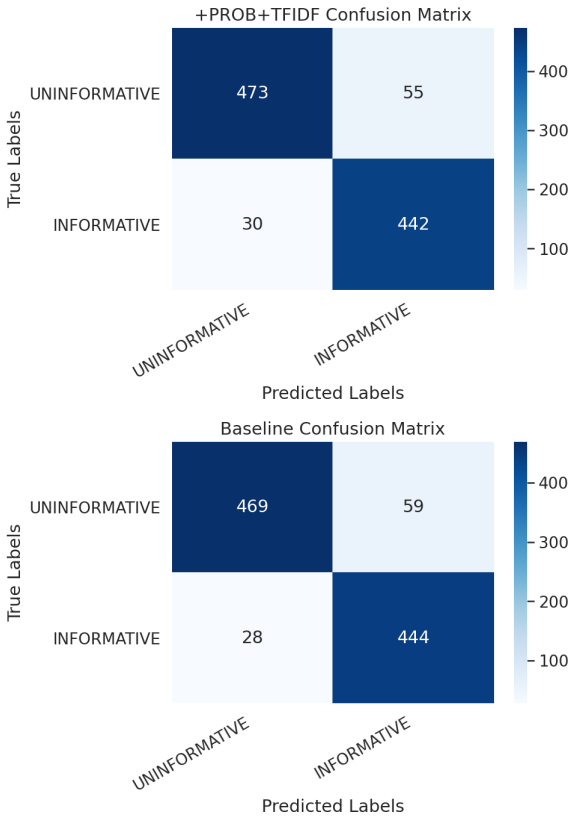

Figure 2: Confusion matrices of the RoBERTa+PROB+TFIDF ensemble model and RoBERTa baseline model respectively.

We additionally explored the intersection of false-positive and false-negative miss-classifications between the two models. We observed that the intersect between the models comprised 87% of the RoBERTa+PROB+TFIDF ensemble model's false-positive miss-classifications, and 77% of the false-negative miss-classifications. We submit that this large intersection within the miss-classifications is indicative of a large proportion of tweets with which the RoBERTa model inherently struggles. Valentini and Masulli (2002) describe how the effectiveness of an ensemble model directly relies upon the accuracy and *diversity* of the individual base learners, and to this regard we submit that more substantial improvements in performance require overcoming the limitations of the RoBERTa model in it's current form.

5 Conclusions and Further Work

We have explored with providing supplementary information to improve the performance of RoBERTa. We have observed that the addition of a handcrafted feature improved performance of a pre-trained bidirectional transformer-based language model, suggesting that for text classification tasks and noisy data sets the inclusion of additional features that distinguish the classes can be beneficial. We intend to explore this concept further, beginning with conducting a statistical test to observe whether the difference between the presence of numeric data in informative tweets is significant.

Additionally, we have experienced success with incorporating TF-IDF features with BERT to a degree, however we present that the size of the document-term matrix, the short length of tweets and domain-specific features of Twitter present potentially a challenge in utilising this optimally.

As the global coronavirus pandemic continues to develop, the nature of what becomes "informative" information will likely develop also. Therefore, in the present task, an approach with a greater level of generalisation is arguably preferable. To this regard, initial testing of parsing the tweets and replacing the terms "coronavirus" and "covid-19" to "coronavirus disease", with a view to leveraging the existing embedding for "disease" to provide contextual information has shown promising results. The application of this new parsing objective presents an opportunity for future work.

Furthermore, we observed in Section 2 that conducting additional *in-domain* pre-training was successfully utilised in relation to pre-training transformer-based models in a similar task. Additionally, in Müller et al. (2020) the authors release a BERT model pre-trained on a data set of tweets in relation to the coronavirus that shows a 10-30% marginal improvement on numerous classification data sets compared to $\text{BERT}_{\text{LARGE}}$. Exploration of these two concepts may provide insight into improving the base RoBERTa model's performance (as discussed in Section 4.1), and present further potential for future work and overcoming the challenges of a noisy data set centred around the topic of coronavirus.

Acknowledgements

We would like to thank the NVIDIA Deep Learning Institute for the provision of AWS credits which we used to access GPU resources in this work.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292.

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. User generated data: Achilles' heel of bert.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019a. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019b. Rumor detection on social media: Datasets, methods and opportunities.

Wah Meng Lim and Harish Tayyar Madabushi. 2020. Uob at semeval-2020 task 12: Boosting bert with corpus level information.

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.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

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*.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

F. P. Shah and V. Patel. 2016. A review on feature selection and feature extraction for text classification. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2264–2268.

Sajad Sotudeh, Tong Xiang, Hao-Ren Yao, Sean MacAvaney, Eugene Yang, Nazli Goharian, and Ophir Frieder. 2020. Guir at semeval-2020 task 12: Domain-tuned contextualized models for offensive language detection.

Giorgio Valentini and Francesco Masulli. 2002. Ensembles of learning machines. In *Neural Nets*, pages 3–20, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP*.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. Uhh-lt at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection.

Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors. 2019. *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020).

A Appendix

Pie Chart Representations of the Intersection between SVM, BERT and RoBERTa