COVCOR20 at WNUT-2020 Task 2: An Attempt to Combine Deep Learning and Expert rules

Ali Hürriyetoğlu
Koç University
ahurriyetoglu@ku.edu.tr

Ali Safaya
Koç University
asafaya@ku.edu.tr

Osman Mutlu
Koç University
omutlu@ku.edu.tr

Nelleke Oostdijk
Radboud University
n.oostdijk@let.ru.nl

Erdem Yörük
Koç University
eryoruk@ku.edu.tr

Abstract

In the scope of WNUT-2020 Task 2, we developed various text classification systems, using deep learning models and one using linguistically informed rules. While both of the deep learning systems outperformed the system using the linguistically informed rules, we found that through the integration of (the output of) the three systems a better performance could be achieved than the standalone performance of each approach in a cross-validation setting. However, on the test data the performance of the integration was slightly lower than our best performing deep learning model. These results hardly indicate any progress in line of integrating machine learning and expert rules driven systems. We expect that the release of the annotation manuals and gold labels of the test data after this workshop will shed light on these perplexing results.

1 Introduction

The COVID-19 pandemic urged various science disciplines to do their best so as to contribute to understanding and relieving its impact. Thus, scholars and practitioners working on information sciences have been dedicating significant effort to help. Collecting and analyzing data published on social media platforms have become the focus in this respect. We joined the community that aims at organizing data collected from social media (in this use case: Twitter), as informative and uninformative. The WNUT-2020 Task 2 considers tweets about recovered, suspected, confirmed and death cases as well as location or travel history of the cases as informative. All other tweets are considered to be uninformative. The organizers did not share an annotation manual nor was a baseline system made available, presumably to prevent use of any other manually annotated data and to encourage broad participation respectively (Nguyen et al., 2020b).[1]

The effort was managed in terms of a shared task, in which the organizers share a dataset that consists of annotated tweets and conduct the evaluation of the submissions. The task requires the participating teams to develop short-text classification systems that facilitate the training and development data to generalize to the test set they release. Although the gold labels of the training and development data were available to the participants, neither the gold labels of the test data nor the annotation guidelines for any part of the data were shared with the participants. Moreover, the test instances were unknown to the participating teams. They were hidden in a larger dataset. Each team was allowed to submit only two outputs of the systems they developed for classifying tweets on the Codalab page of the task.[2] The highest score in terms of F1 positive class of each team was used to rank them in the leaderboard.

Integrating automatically created machine learning based (ML) models with manually formulated rules to tackle a text classification task promises the best of both worlds. We pursued this goal by integrating the output of two deep learning models and a rule-based system under the team name COVCOR20. Although the integration slightly improves the total performance on the training and development sets in a cross-validation setting, the overall performance on the test data turned out to be slightly worse than our best ML system. Our best submission was ranked 22nd among 55 teams. The integration of our systems would be ranked 27th if its score were used as the final score for our team.

The deep learning models and the rule-based system are introduced in Sections 2 and 3 respectively. Next, the Section 4 describes how we integrate the output of these systems. Then Section 5 provide the results and their discussion. Finally, we conclude this report and share our future plans continuing in this line of research in Section 6.

2 Deep learning models

We created various deep learning models by merging the training and development data released by the organizers. We applied k-fold cross-validation by splitting the data into five parts. The average performance in terms of F1 of five optimization and test iterations is used to compare these models. The hyper-parameter optimization is focused on F1 of the positive class. Table 1 provides performances of the models we tested. CNNText, BiLSTM, and BERT-CNN were borrowed from Safaya et al. (2020), BERT is the standard model from Devlin et al. (2019), and BERTweet is the model created by Nguyen et al. (2020a).

We developed a novel model that we call Fused-1 and Fused-2 that has a modified version of Char-RNN[3] is a character language model that consists of two LSTM layers and is pretrained on 112 thousand tweets related to COVID19 which we collected. In Fused-1 CharRNN is applied in parallel with BERT, while in Fused-2 it is used in parallel with BERTweet. Just as in BERT-CNN, we feed the output of the last hidden layer of Char-RNN to a Convolutional Neural Network (CNN). We also have another CNN that takes BERT or BERTweet output as its input. Upon this, we concatenate the output of these two CNNs (the first one using BERT or BERTweet as embedder and the latter using CharRNN as embedder), and we feed it to a fully connected layer to get the final output[4].

We used Fused-2 as it is the best performing model in the aforementioned cross-validation setting. Moreover, we observed that the average number of the sentences in wrongly predicted instances is larger than the correctly predicted cases using the best performing deep learning model. Therefore, we generated a second output for each tweet by

Model	F1	Std
CNNText	.8543	.0116
BiLSTM	.8595	.0089
BERT	.9432	.0066
BERT-CNN	.9478	.0060
BERTweet	.9499	.0029
Fused-1	.9552	.0022
Fused-2	.9560	.0024

Table 1: Model performances in terms of F1 and Standard Deviation (Std).

predicting each sentence separately. If at least one sentence in a tweet is predicted as informative, the prediction for the whole tweet becomes positive, i.e. informative. The precision decreased, the recall increased, and the F1 slightly decreased in this setting.

3 Rule-based system

The rule-based system uses a set of handcrafted rules constructed by an expert for the task at hand and a task-specific lexicon. The system is designed to identify only instances of the positive class, which for this particular task are the tweets that are considered to be informative. The rules describe, in terms of linguistic patterns, the salient part(s) of these tweets. Whenever a rule fires, the tweet will be labelled as informative; tweets for which no rule is found to apply are considered non-informative by default. A similar rule-based approach has been used successfully in other shared tasks (e.g. Hürriyetoğlu and Oostdijk (2017); Oostdijk and Hürriyetoğlu (2017); Oostdijk and van Halteren (2019)). There, the approach has been shown to yield consistently high precision across different datasets, while recall generally falls short of that yielded by ML approaches.

In this case the lexicon we compiled has 779 entries. Entries here are of the form: word type – semantic/syntactic word class – (optionally) label, e.g.

case N informative	death N informative
first NUMord	covid19 Ncorona
new ADJ	confirmed ADJ

In order to have maximum control over the strings matched, in most cases entries are delimited by word boundary markers (\b), indicating the beginning and the end of a word or multi-word expression. An exception was made with the entries referring to the virus (including for example,

---

covid19, covid-19, coronavirus) where not including the initial word boundary marker allows for the matching of instances where the word is part of a hashtag.

The rules set comprises 409 rules. They describe how the lexical entries can combine to form larger strings, the length of which in the present case varies between 2 to 7 words. Example rules are[5]

    NUMord *N
    ADJ Ncorona *N
    NUMord ADJ *N

which account for instances like *first case*, *first confirmed death*, and *new covid19 deaths*. One of the strengths of the rule-based approach is that the rules generalize beyond the instances observed in the training set, provided that the individual words appear in the lexicon.

The performance achieved with the rule-based system was as follows: on the training set precision was .845, recall .893 and F1 86.9; on the validation set precision was .77, recall .852 and F1 .809. We are able increase the precision at the expense of decreasing recall by requiring the number of detected spans to be equal or more than a certain threshold. For instance, the precision and recall become .91 and .53 in average if we set this threshold to be 2.

4 Integrating outputs

We have shown that ML and rule-based approaches yield complementary output for text classification tasks in various settings in our previous work (Hürriyetoğlu and Oostdijk, 2017; Oostdijk and Hürriyetoğlu, 2017; Oostdijk and van Halteren, 2019). Therefore, we continued to work along this line by focusing on integrating the outputs of these systems in the scope of this shared task. However, voting and applying OR and AND operations on the predictions did not outperform the best deep learning model Fused-2. The only integration that perform slightly better was giving precedence to the rule-based system in case the two output versions of Fused-2 conflict and the rule-based system predicts a tweet as uninformative. In case the rule-based system predicts it informative, we used prediction of the best system, which is Fused-2 on all text of a tweet. This adjustment increased F1 from .9560 to .9581 in our cross-validation setting.

5 Results

We submitted the best performing deep learning model, which is Fused-2 using all text of a tweet and the best performing integration. The deep learning model yielded .8887 and the integration yielded 0.8856 F1. Our team ranked 22 with these scores.[6] The best overall F1 score in the shared task was .9096.

The decrease of the scores from around .9560 on development set to .8887 on the test set shows that rule-based system could have had a chance by itself. However, the restriction of the submission count to be 2 did not allow us to evaluate this.

6 Conclusion and future work

We have presented our effort in the scope of a shared task that aims at pushing the state of the art for classifying short-texts (i.e. tweets in the reported use case), as informative or uninformative.

We could extend the training set with cluster mining using Relevancer (Hürriyetoğlu et al., 2016), use the rule-based system to extend the training set (Hürriyetoğlu, 2019), or use the rule-based system to generate fine-grained data that can be used in a multi-task setting.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ali Hürriyetoğlu. 2019. *Extracting Actionable Information from Microtexts*. Ph.D. thesis, Radboud University Nijmegen.

Ali Hürriyetoğlu, Christian Gudehus, Nelleke Oostdijk, and Antal van den Bosch. 2016. *Relevancer: Finding and Labeling Relevant Information in Tweet Col-*

---

[5]The asterisk indicates the element from which the string inherits its label.

[6]All results of the competition are on: `https://competitions.codalab.org/competitions/25845#results`

lections, pages 210–224. Springer International Publishing, Cham.

Ali Hürriyetoğlu and Nelleke Oostdijk. 2017. Extracting Humanitarian Information from Tweets. In *Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness*, Aberdeen, United Kingdom.

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*.

Nelleke Oostdijk and Hans van Halteren. 2019. Team taurus at semeval-2019 task 9: Expert-informed pattern recognition for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1247–1253.

Nelleke Oostdijk and Ali Hürriyetoğlu. 2017. Detecting the need for resources and their availability. In *Majumder, P.; Mitra, M.; Mehta, P.(ed.), FIRE 2017 Working Notes. Working notes of FIRE 2017–Forum for Information Retrieval Evaluation*, pages 34–37. Bangalore, India:[Sn].

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media.