A fuzzy deep learning approach to health-related text classification

Nasser Ghadiri[1][0000-0002-6519-6548], Ali Ghadiri[1] and Afrooz Sheikholeslami[1]

[1] Department of Electrical and Computer Engineering, Isfahan University of Technology,
Isfahan 84156-83111, Iran
{nghadiri@iut.ac.ir, alighadiri@ec.iut.ac.ir,
afrooz.sheikholeslami@ec.iut.ac.ir}

Abstract. Following the tremendous amounts of text generated in social networks and news channels, and gaining valuable and dependable insights from diverse sources of information is a tedious task. The challenge is increased during specific periods, for example, in a pandemic event like Covid-19. Existing text categorization methods, such as sentiment classification, aim to help people tackle this challenge by categorizing and summarizing the text content. However, the inherent uncertainty of user-generated text limits their efficiency. This paper proposes a novel architecture based on fuzzy inference and deep learning for sentiment classification that overcomes this limitation. We evaluate the proposed method by applying it to well-known health-related text datasets and comparing the accuracy with state-of-the-art methods. The results show that the proposed fuzzy fusion methods increase the accuracy compared to individual pretrained models. The model also provides an expressive architecture for health news classification.

Keywords: Text Categorization, Sentiment Analysis, Fuzzy Inference, Text Mining, Deep Learning.

1 Introduction

During pandemic events, people generally use the various web and social media content and use them to communicate with society and express ideas and news [1]. Social media's great potential could help people be informed about the COVID-19 spread, knowing about the spread patterns, and preventing the risks [2].

A challenging task for the citizens is reading, analyzing, and interpreting the massive amount of news and related content, especially in pandemic events. Following the large volume of text generated in social networks and news channels and gaining valuable and dependable insights from diverse information sources is a tedious task. The challenge is increased during specific periods, for example, in a pandemic event like the outbreak of Covid-19 [1].

Text classification methods, such as sentiment analysis, aim to help users tackle this challenge by categorizing and summarizing the text content. Earlier sentiment analysis methods were based on statistics and machine learning, while newer methods often provide higher accuracy by pretraining deep neural network models [3]. One of the most common pretrained models was BERT, proposed by Google [4]. A deep neural network is pre-trained by extracting information from very large text corpora to adjust the network weights in these methods. The network then is fine-tuned for specific downstream tasks like sentiment analysis. The pretraining often provides higher accuracy compared to primary methods. An optimized model based on BERT is RoBERTa [5], developed by Facebook containing more training parameters and performs well on different tasks, including next sentence prediction and question answering. It handles long sentences more efficiently. TwitterBERT [6] is also proposed for processing tweets, COVID-Twitter-BERT [7], or CT-BERT tailored for tweets related to Covid19.

However, the inherent uncertainty of user-generated text limits the efficiency of PLM methods. For instance, a Covid19-related tweet could be better processed by CT-BERT, but the RoBERTa model may also provide high accuracy for some longer tweets in this domain. While a specific tweet belongs to the set of Covid-related tweets, it may also be a member of the set of long tweets to some degree. Therefore, we need to cope with the fuzziness in using PLMs like BERT to analyze the tweets.

In this paper, we propose a hybrid architecture based on three different pretrained models to cover different tweets. Then it uses two fuzzy methods for the fusion of data from the pretrained BERT-based models. The first fusion is based on the Choquet fuzzy integral that has shown remarkable performance in the fusion of data for decision-making scenarios [7]. The second is designed as a fuzzy rule-based model for the fusion of classifiers [8]. The details of the proposed model will be described in the next section.

2 Proposed Model

The proposed model is a hybrid of three state-of-the-art pre-trained models fused by two fuzzy fusion methods. The overall process is illustrated in Fig. 1. The first step is pre-processing of the input data. Raw twitter data contains much noise and is unstructured and informal. Therefore, pre-processing on Twitter data will help the pre-trained models in better performance. The detailed steps of pre-processing are described in Section 3. After pre-processing, every tweet is passed to three different pretrained models for classification. The pretrained model will be described in Section 2.1. In the third step, the results are fused using two fuzzy fusion methods. The first fusion method is the Choquet integral that uses validation data from the pretrained models. The second is a fuzzy rule-based fusion module that uses the classification output from the pre-trained models and meta-data for tuning the rules. More details about the modules will be presented as follows.

2.1 The pretrained models

The pre-processed input tweets are fed into different pretrained models. We selected three models, including BERT, RoBERTa, and Covid-Twitter-BERT, explained below.

BERT base uncased. BERT is a transformers model which is pretrained on the BookCorpus dataset that consists of about 11K unpublished books and content from English Wikipedia. It will function as a general classifier in our model.
RoBERTa Large. RoBERTa is based on BERT and extended its dataset using CC-News, OpenWebText, and Stories, which improve its performance. Furthermore, RoBERTa is not pretrained with the next-sentence prediction (NSP) objective. It is only pretrained with the Masked language modeling (MLM) objective and uses a larger batch size than BERT. This model is expected to classify long tweets more accurately.
COVID-Twitter-BERT (CT-BERT). The architecture of COVID-Twitter-BERT is similar to BERT. Unlike BERT, which is pretrained on the English Wikipedia dataset and may not have a deep understanding of texts related to COVID-19, CT-BERT is optimized for using texts related to COVID-19.

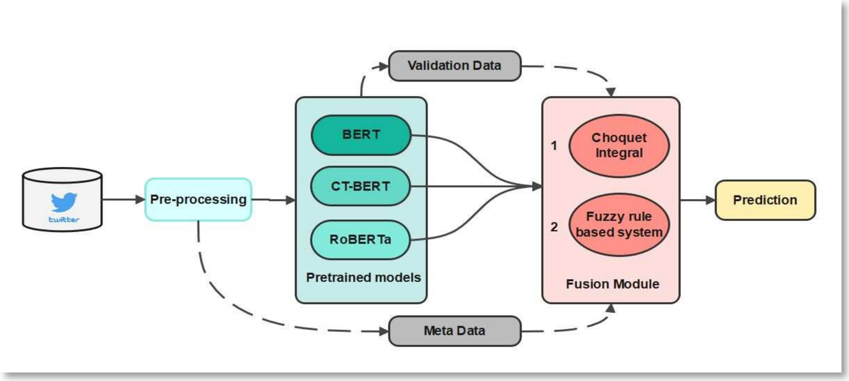


Fig. 1. The overall process for the sentiment analysis of Covid19 related tweets.

2.2 Fuzzy Choquet integral fusion

We used two methods for the fusion of the results obtained from the pretrained models. The first fusion method is based on fuzzy Choquet integral [7]. Given $x_1, x_2, \ldots, x_n$ as a set of criteria, and v as the fuzzy measure, the Choquet integral is given by:

$$C_v(x) = \sum_{i=1}^{n} [x_i - x_{i-1}] v(H_i)$$

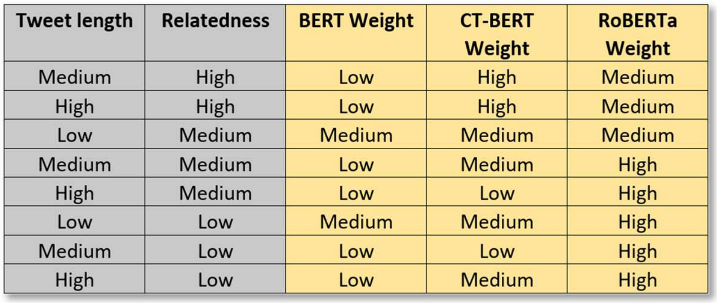
Where $H_i = \{i, \ldots, n\}$ is the subsets of indexes of the $n - i + 1$ most significant components of x [9]. We used the fuzzy Choquet integral in two settings; (a) fuzzy fusion using Choquet integral trained on *densities*, and (b) fuzzy fusion using Choquet integral trained on *validation scores*. The results of the fuzzy fusion of pretrained models using both techniques will be reported in Section 3.

2.3 Fuzzy rule-based fusion

Our second fuzzy fusion method for combining the results from the pretrained classifiers is a fuzzy rule-based system [8]. It has two inputs that contain metadata, and three outputs to set the weights of three classifiers. The fuzzy rules are shown in Fig. 2. The two metadata inputs are described as follows.

Meta Data. For enriching the fuzzy rule-based fusion of pretrained models, we propose using two meta-parameters extracted from every input tweet. The first parameter is the *tweet length*. It could be observed that some pretrained models like RoBERTa perform well for longer texts. So our fusion method takes advantage of tweet-length information to tune the weights of the pretrained model outputs accordingly.

The second parameter is the *relatedness* of the tweet to the Covid19 domain. Some pretrained models like CT-BERT are targeted at tweets that are highly related to this domain. So our fusion method computes the similarity between every tweet with a fixed set of Covid19 keywords using the Jaccard similarity measure. Then the rule-based model takes advantage of tweet relatedness information to tune the weights of the pretrained model outputs accordingly.



Tweet length	Relatedness	BERT Weight	CT-BERT Weight	RoBERTa Weight
Medium	High	Low	High	Medium
High	High	Low	High	Medium
Low	Medium	Medium	Medium	Medium
Medium	Medium	Low	Medium	High
High	Medium	Low	Low	High
Low	Low	Medium	Medium	High
Medium	Low	Low	Low	High
High	Low	Low	Medium	High

Fig. 2. The fuzzy rules for the fusion of pretraining outputs.

3 Experimental Evaluation

In this section, we present the evaluation of our proposed method. We will describe four steps for data collection, pre-processing, classification using three different pretrained models, and the fuzzy fusion of the classification results. We used the Python language and Google Collab for running our experiments. For all experiments, the batch size was 16, and the sequence maximum length was 40. We concatenated the validation and the training data and performed a 5-fold cross validation to train our models. Each fold is trained in 5 epochs using early stopping with patience of 3.

3.1 Data Collection

We have collected data from the Lopezbec repository[1] , which contains an ongoing collection of tweets associated with the novel coronavirus COVID-19 since January 22nd, 2020. Only tweets in English were collected from July 2020 to January 2021 and among them, those which met the criteria for like, retweet, and sequence length were selected. Accordingly, we collected 55,785 tweets which are categorized as positive, negative, and neutral. In the next step, we used TWARC to hydrate the tweet-IDs. A sample of the dataset is shown in Table 1.

Table 1. A sample of tweets about Covid19.

Tweet Text	Sentiment Label
Congrats to senior scientist Alex MacKenzie, winner of the 2020 @CHEO Research Institute Osmond Impact Award. Most recently famed for his #COVID19 research, his decades of work on #RareDisease, spinal muscular atrophy and #diabetes has been cited over 10,000 times. #CDNhealth https://t.co/gLSVQ7pFBY	Positive
@woodenfam1 Not canceled, just pushed back. The pandemic has caused an immense change in our production schedule for all products. N scale James, HO Peter Sam, HO Daisy, and HO Oil Tanker Troublesome Truck #6 all were not able to make it in 2020	Negative
This morning I was asked to address recent comments from the White House about Florida's vaccine allocation. Here is my statement: https://t.co/imwvT67sRo	Neutral

As discussed earlier in Section 3, and the contents of sample tweets suggest, pre-processing of the tweets is essential before the classification step. In the next part, we describe the pre-processing tasks.

3.2 Data Preprocessing

The following series of techniques are applied in the given order to improve the text.

1. People highly use hashtags in social media to represent topics, i.e., #COVID-19, #StayHome, #StaySafe, and #Coronavirus. We performed the cleaning of the text by removing hashtag characters and segment the hashtag texts using wordninja package.
2. Convert tweets to lower cases: The tweets may be in lower or upper cases. In this step, we convert all tweets into lower case.
3. Unscape HTML tags and eliminate hyper-links, @mentions , emails, and numbers.

[1] https://github.com/lopezbec/COVID19_Tweets_Dataset#data-collection-process-inconsistencies

4. In some tweets, shortened versions of words (contractions) exist that should be converted by removing specific letters. For instance, "don't" will convert to "do not" and "what's" will be "what is". Converting each contraction to its expanded helps with text standardization.
5. We eliminated punctuation, and special characters from the dataset as these do not help detect sentiment. Furthermore, we removed emojis using the python emoji library to *demojise* the emojis and replace them with a short textual description.
6. Removing stop words is a common method to reduce the noise in textual data. Removing stop words does not affect understanding a sentence's sentiment valence. We removed stop words using the stop-word library in NLTK. This library includes 179 stop words of the English language.
7. Lemmatization: The purpose of this step is to convert words to their root to get distinguishing words. Among the methods to implement this step, we used the spaCy package that efficiently addresses the challenges that we face for this step.

3.3 Experimental Results

The results of accuracy provided by each classifier through different epochs are shown in Fig. 3. It could be observed that the accuracy of the BERT model declines after three epochs. However, BERT has shown the lowest variation in accuracy compared to other models. The RoBERTa pretrained classifier shows high variations but converges at the fifth epoch. The results show that the CT-BERT model provides more stable results as expected from a pretrained model that is specifically tailored to tweet inputs.

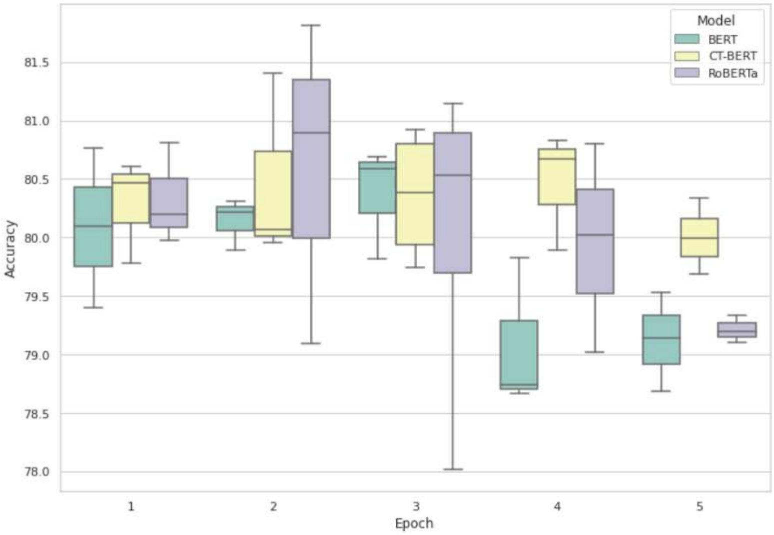


Fig. 3. The accuracy ranges of different pretraining models.

The mean accuracy values of individual pretrained models for validation data are shown in Table 2. On average, the CT-BERT model is more accurate than other methods in the training phase. However, as we observe in Fig. 3, no single model could perform more accurately than other models during the training. Moreover, the accuracy for test phase depends on the generalizability of the model and, often less than the accuracy for validation data. Therefore, we transfer the results of the pretrained models to the fuzzy fusion module for potentially improved accuracy.

Table 2. The accuracy of different classifiers for *validation* data

Model	Accuracy
BERT base	88.2
RoBERTa	91.3
CT-BERT	92.0

The accuracy of different methods for test data is shown in Table 3. We can observe that both of our proposed fusion methods (Choquet integral with two settings and the rule-based method) provide higher accuracy compared to individual BERT, RoBERTa, and CT-BERT methods.

Table 3. The accuracy of different classifiers for *test* data

Model	Accuracy
BERT base	80.3
RoBERTa	81.1
CT-BERT	81.0
Fuzzy fusion 1 - CHI trained on validation scores	**81.4**
Fuzzy fusion 1 - CHI trained on densities	**81.6**
Fuzzy fusion 2 - Fuzzy rule-based system	**82.0**

The rule-based fusion has performed slightly better than Choquet integral. This could be the result of feeding more information to the rule-based module through metadata about tweet length and the degree of relatedness to Covid19 tweets.

4 Conclusion and Future Work

This research aimed to help the citizens gain more information from the massive amount of content generated during pandemic events. Although many state-of-the-art pretrained language models are developed to classify such content, the accuracy of the models is limited due to the inherent uncertainty of user-generated content. We proposed two fuzzy fusion approaches to get a higher accuracy by combining the different pretrained models. Both of the fuzzy Choquet integral and fuzzy rule-based fusion methods performed better than individual models. The meta-data extracted from the input tweets also contributed to the increased accuracy of the rule-based model.

Future work may focus on improving the rule-based fusion by learning the fuzzy sets for both input and output variables. The metadata that is fed to the rule-based module may also be enriched by adding more features from the input tweets. The Choquet integral will also be extended if more pretrained classifiers become available in the future. In this case, the Choquet integral would be expected to perform better than the rule-based fusion that requires adding more output variables for new pretrained modules.

References

1. Alshamrani, S., Abusnaina, A., Abuhamad, M., Lee, A., Nyang, D., Mohaisen, D.: An Analysis of Users Engagement on Twitter During the COVID-19 Pandemic: Topical Trends and Sentiments. Presented at the December 11 (2020). https://doi.org/10.1007/978-3-030-66046-8_7.
2. Rashid, M.T., Wang, D.: CovidSens: a vision on reliable social sensing for COVID-19. Artif. Intell. Rev. 54, 1–25 (2021). https://doi.org/10.1007/s10462-020-09852-3.
3. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep Learning Based Text Classification: A Comprehensive Review. arXiv. 1, 1–43 (2020).
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1, 4171–4186 (2019).
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach, http://arxiv.org/abs/1907.11692, (2019).
6. Azzouza, N., Akli-Astouati, K., Ibrahim, R.: Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. In: Advances in Intelligent Systems and Computing. pp. 428–437. Springer (2020). https://doi.org/10.1007/978-3-030-33582-3_41.
7. Choquet, G.: Theory of capacities. Ann. l'institut Fourier. 5, 131–295 (1954). https://doi.org/10.5802/aif.53.
8. Wang, D., Keller, J.M., Andrew Carson, C., McAdoo-Edwards, K.K., Bailey, C.W.: Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion, (1998). https://doi.org/10.1109/3477.704297.
9. Zhao, Y., Cen, Y.: Data Mining Applications with R. Elsevier Inc. (2013). https://doi.org/10.1016/C2012-0-00333-X.