Fake News and Hostile Posts Detection Using an Ensemble Learning Model

Siyao Zhou, Jie Li, and Haiyan Ding(✉)

Yunnan University, Yunnan, People's Republic of China

Abstract. This paper describes the system submitted to Constraint 2021. The purpose of this task is to identify fake news in English and hostile posts in Hindi. We experimented with the pre-trained model based on the transformer and adopted the method of Ensemble Learning. We observed that the model ensemble was able to obtain better text classification results than a single model, the weighted fine-grained F1 score of our model in subtask B was 0.643998 (ranking 1/45).

Keywords: Fake news · Hostile posts · Constraint 2021 · Pre-trained model · Transformer · Ensemble learning

1 Introduction

With the increasing popularity of the Internet, the use of social media has grown rapidly in the past few years and has become a great platform for people living far away to communicate. Many people posts their opinions, thoughts, and comments on social networking sites such as Facebook, Twitter, etc. This has also led to the spread of hate speech and fake news on the Internet. Cyber hatred can not only affect one's mental health, but also turn into violence in the real world, so this issue needs attention.

Constraint 2021 [1] encourages interdisciplinary researchers to work on multilingual social media analytics by providing a platform to test hostile posts and fake news detection through organized competitions. The challenge collects data from Twitter and Facebook and provides two subtasks, COVID19 Fake News Detection in English, which focuses on detecting Fake News in English related to COVID19. The other subtask is a hostile posts detection in Hindi, with a valid set of categories including false news, hate speech, offensive, defamatory and non-hostile speech. This subtask is relatively more challenging than the first one because not only the number of classes is increased, but also it is a multi-category classification problem with multiple tags.

To solve this problem, we used the pre-training model BERT and Ensemble Learning to accomplish these two tasks. Compared with other methods, it relies less on preprocessing and feature engineering, and the model has been proved to be very effective in natural language processing tasks across multiple languages.

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 74–82, 2021.
https://doi.org/10.1007/978-3-030-73696-5_8

The rest of the paper is organized as follows: Sect. 2 reviews the related work. Sections 3 and 4 respectively describe the relevant data and model approaches we use. We discuss our experiment in Sect. 6, which describes our results. Finally, Sect. 7 summarizes our work and discusses further improvements.

2 Related Work

As social media has become more popular over the years, hostile posts have become more common on these platforms. Hostile posts detection is a broad area of research that attracts many people. Here we briefly describe some of the work that has been done in this regard. Machine learning and natural language processing have made breakthroughs in detecting hostile posts on online platforms. Much scientific research has focused on using machine learning and deep learning methods to automatically detect fake news and hostile posts.

Some studies have shown that the deep learning model with embedded words can achieve better results in text classification tasks. Waseemc [2] used SVM and LR classifiers to detect racist or sexist content and tested the impact of hate speech knowledge on the classification model. Thomas et al. [3] used logistic regression, Naive Bayes, decision tree, random forest, and linear SVM models for automatic hate speech detection. After many years of research, RNN [4] model has achieved good results in emotion analysis tasks. The latest trend in deep learning has led to better sentence expression.

Recent methods used semantic vectors such as Word2vec [5] and GloVe [6] to better represent words and sentences. These methods are superior to the earlier BOW method because similar words are closer together in potential space. As a result, these continuous and dense representations replace earlier binary features, leading to the more efficient encoding of input data. Kai Shu [7] proposed a tri-relationship embedding framework TriFN, which models publisher-news relations and user-news interactions simultaneously for fake news classification.

In recent years, transformer [8] based language model can be used for pre-training with specific targets on a large corpus to obtain rich semantic features of the text. BERT(Bidirectional Encoder Representations from Transformers) [9] model further increases the generalization ability of the word vector model, and fully describes character-level, word-level, sentence-level, and even inter-sentence relationship characteristics. The ensemble learning [10] is considered the most advanced solution to many machine learning challenges. These methods improve the prediction performance of a single model by training multiple models and combining their prediction results.

3 Datasets

The task data set is provided by Constraint 2021 organizer, and we present the Constraint data set statistics in the Table 1. In the task of detecting COVID-19 fake news, data [11] were collected from Twitter, Facebook, Instagram, and other social media platforms, as shown in Table 1. For these given social media posts,

what we need to accomplish is a binary categorization task with two different form categories:

- **fake**: This class contains posts that are untrue or contain error messages. Example: If you take Crocin Thrice a day you are safe.
- **true**: This class contains posts that are logical and realistic or that contain real information. Example: Wearing mask can protect you from the virus.

Table 1. Statistics of the English Sub-task A set provided by the organizers.

Sub-task A	Real	Fake	Total
Train	3360	3060	6420
Valid	1120	1020	2140

Table 2. Statistics of the Handi Sub-task B train set and valid set provided by the organizers.

Train		Valid	
Label	Number	Label	Number
non-hostile	3050	non-hostile	435
fake	1009	fake	144
hate	478	hate	68
offensive	405	offensive	57
defamation	305	defamation	43
hate,offensive	163	hate,offensive	23
defamation,offensive	81	defamation,offensive	11
defamation,hate	74	defamation,hate	10
defamation,fake	34	fake,offensive	4
defamation,hate,offensive	28	defamation,hate,offensive	4
fake,offensive	8	defamation,fake	4
fake,hate	27	defamation,fake,offensive	3
defamation,fake,offensive	24	fake,hate	3
defamation,fake,hate	9	defamation,fake,hate	1
defamation,fake,hate,offensive	9	defamation,fake,hate,offensive	1
fake,hate,offensive	4		

For hostile posts in Hindi, coarse-grained assessment is a dualistic categorical task, divided into hostile and non-hostile. In a fine-grained assessment, it is a

multi-label multi-category classification problem, where each posts can belong to one or more of these rival categories. The relevant set of valid categories includes false news, hate speech, offensive, defamatory and non-hostile speech, as shown in the following forms:

- **fake**: A claim or information that is verified to be not true.
- **hate**: A posts targeting a specific group of people based on their ethnicity, religious beliefs, geographical belonging, race, etc., with malicious intentions of spreading hate or encouraging violence.
- **offensive**: A posts containing profanity, impolite, rude, or vulgar language to insult a targeted individual or group.
- **defamation**: A mis-information regarding an individual or group.
- **non-hostile**: A posts without any hostility.

In Table 2, we listed the specific number of posts for the training set and the valid set [12].

4 Methodology

4.1 BERT

The transformer-based language model has received a lot of attention in the past, where BERT has worked well for many natural language processing tasks. The model structure is shown in Fig. 1. Given a sentence or a paragraph as input, the input sequence adds a [CLS] token at the beginning of the sentence, and the [SEP] token serves as a separator between the sentences or a marker at the end of the sentence. Then each word in the input sequence is converted into its corresponding word vector, and the position vector of each word is added to reflect the position of the word in the sequence.

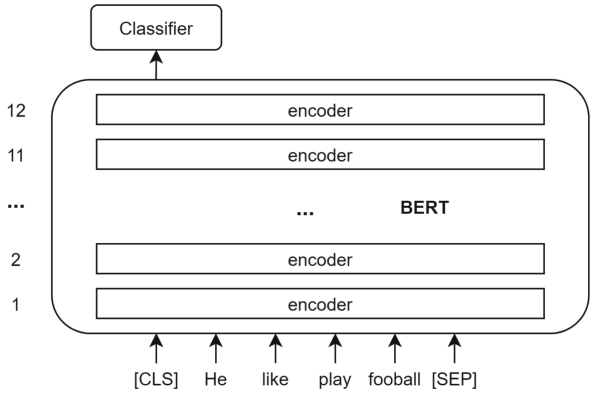

Fig. 1. Model BERT

These word vectors are then inputted into a multi-layer Transformer network, and the relationship between words is learned through the self-attention mechanism to encode their context information. Then a feedforward network is used to output the vector representation of each word that integrates the context characteristics through nonlinear changes. Each encoder layer is mainly composed of two sub-layers: the multi-head self-attention layer (multi-head self-attention mechanism) and the feedforward network layer.

Multi-head self-attention will calculate several different self-attention parameters in parallel, and the results of each self-attention will be spliced as the input of the subsequent network. After that, we get the representation of the words that contain the current context information, which the network then inputs to the feedforward network layer to calculate the characteristics of the nonlinear level.

In each layer of the network, the residual connection introduces the vector before the self-attention mechanism or the feed-forward neural network to enhance the output vector of the self-attention mechanism or the feed-forward network. It also uses the normalization method that maps multi-dimensional vectors of nodes of the same layer into an interval so that the vectors of each layer are in an interval. These two operations are added to each sublayer to train the deep network more smoothly. After the text context features are extracted, they are input to the classifier.

4.2 Ensemble

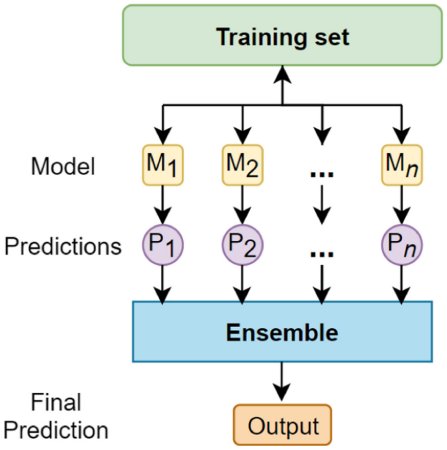

Fig. 2. Ensemble Learning

In the supervised learning algorithm of machine learning, our goal is to learn a stable model that performs well in all aspects, but the actual situation is

often not so ideal, sometimes we can only get multiple models that perform well in some aspects. To mitigate this, ensemble learning can be used to reduce overfitting and improve model generalization. Ensemble learning is the combination of several weak supervised models to get a better and more comprehensive strong supervised model. The underlying idea of ensemble learning is that even if one weak classifier gets a wrong prediction, other weak classifiers can correct the error back. Therefore, ensemble learning is widely used to combine multiple fine-tuning models, and the ensemble BERT model is often more effective than a single BERT model, the model structure is shown in Fig. 2.

Our method uses Stratified 5-fold cross-validation to generate different training data sets, and then get multiple basic classifiers based on these training data sets respectively. Finally, we combine the classification results of basic classifiers to get a relatively better prediction model. We use hard voting to determine the final category, aggregate the categories predicted by each classifier, and then select the category with the most votes. The output of the ensemble model is the prediction with the highest probability. This voting classifier can often be more accurate than a single optimal classifier.

5 Experiment

To enable the model to learn the appropriate semantic characteristics, we consider cleaning up the noise in the data set to obtain clean data. We use the NLTK library for the English and Hindi raw data sets to perform the specified preprocessing tasks.

First, we remove the string that starts with the @ symbol, because the string represents the user's name, it does not contain an expression, and it degrades the model's performance. After that, we remove tags, punctuation, URLs, and numbers, because strings usually start with https:// and have no semantics and need to be removed before further analysis. So it's considered noisy data. We eventually convert emoji into language expressions to produce both pure English and pure Hindi texts in tweets.

We use a BERT-based model from the Huggingface[1] library as our pre-trained language model. The HuggingFace Transformers package is a Python library that provides pre-trained and configurable models for a variety of NLP tasks. It contains pre-trained BERT and other models suitable for downstream tasks. To accomplish this task, we set up five bert-base-uncased models for ensemble learning, the classifier is a linear layer of 768×5 dimensions, the random seed is set to 42. All of our fine-tuning models in the 2 subtasks were trained using the Adam optimizer and CrossEntropy Loss. The learning rate is 2e-5. The epoch and the maximum sentence are set as 3 and 128 respectively. The batch size is set to 32, and the gradient step size is set to 4, as shown in Table 3.

The output of the model is mapped from 0 to 1 by the activation function sigmoid. For sub-task B, the threshold is set to 0.2 to classify the output. For

[1] https://huggingface.co/models.

Table 3. Experimental parameters

Hyperparameters			
learning_rate	2e−5	gradient_accumulation_steps	4
max_seq_length	128	warmup_rate	0.1
batch_size	8	dropout	0.1
attention_dropout	0.1	epoch	3
random_seeds	42		

the predicted value of each label, when it reaches 0.2, it is determined that the label exists, and when it is less than 0.2, it is determined that the label does not exist.

6 Results

The results obtained through the evaluation of valid data will be submitted to the organizers of the shared task for final competition evaluation. Based on the test data, they evaluate each file submitted by all participating teams for each subtask. The final ranking of all teams is determined by the Weighted Average F1 Score, we perform a comparative test based on the evaluation documents provided by the organizer, as shown in Table 4 and Table 5.

Table 4. Prediction results under different methods of subtask A

Method	Accuracy	Precision	Recall	F1-score
CNN	0.8260	0.7935	0.8153	0.8355
LSTM	0.8790	0.8992	0.8960	0.8761
BERT	0.9700	0.9701	0.9700	0.9701
Ensemble_CNN	0.8881	0.9010	0.8963	0.8864
Ensemble_LSTM	0.8920	0.9119	0.9010	0.8896
Ensemble_BERT	0.9766	0.9766	0.9766	**0.9766**

At present, the pre-training models based on deep learning are better than the former CNN and LSTM. The integrated BERT model we used is an improvement in the classification task over the previous approach.

As we can see, our model integration method is better than the single model on the weighted average F1 score, especially for sub-task B, the weighted fine-grained F1 Score improved by 0.12. At the same time, our method is superior to the previous single model in the evaluation of fine granularity in the multi-label classification task, with an improvement of about 0.1–0.2. In addition, we can see that subtask B gets a lower F1 score than subtask A. This may be mainly

Table 5. Prediction results under different methods of subtask B

Method	Coarse grained F1	Defamation F1 score	Fake F1 score	Hate F1	Offensive F1	Fine grained F1
CNN	0.771034	0.272731	0.635026	0.412833	0.548388	0.498829
LSTM	0.813689	0.379535	0.618981	0.512246	0.557526	0.533612
BERT	0.933503	0.319489	0.748872	0.452991	0.582931	0.56253
Ensemble_CNN	0.860298	0.275692	0.766347	0.506428	0.568225	0.576851
Ensemble_LSTM	0.889621	0.355708	0.789204	0.486598	0.614536	0.598620
Ensemble_BERT	**0.960679**	0.455172	0.812214	0.591045	0.589744	**0.643998**

due to the imbalance of the subtask B data set, with differences between the five categories.

The results reported by the organizers showed that the competition among the participating teams was very intense, and our best performance in English subtask A was 0.9766 macro F1 score, with the proposed method achieving a difference of 0.01 from the best result. In Hindi subtask B, the coarse-grained F1 score was 0.96, ranking 8th, and the weighted fine-grained F1 score was 0.64, ranking 1st.

7 Conclusion

In this paper, we used the pre-trained language model BERT to classify hate and offensive content in social media posts. Based on the BERT model, we also adopted the method of ensemble learning. The experiment verified the practicability and effectiveness of this method, and the research results provide a solid foundation for the further study of multilingual hate speech.

References

1. Patwa, P., et al.: Overview of constraint 2021 shared tasks: detecting English COVID-19 fake news and Hindi hostile posts. In: Chakraborty, T., et al. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 42–53. Springer, Cham (2021)
2. Waseem, Z.: Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142 (2016)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009, 2017
4. Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K., Soman, K.P.: Stock price prediction using LSTM, RNN and CNN-sliding window model. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1643–1647. IEEE (2017)
5. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al'.s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)

6. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
7. Shu, K., Wang, S., Liu, H.: Beyond news contents: the role of social context for fake news detection. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 312–320 (2019)
8. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **8**(4) (2018)
11. Patwa, P., et al.: Fighting an infodemic: COVID-19 fake news dataset. arXiv preprint arXiv:2011.03327 (2020)
12. Bhardwaj, M., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Hostility detection dataset in Hindi. arXiv preprint arXiv:2011.03588 (2020)