# Trust or Suspect? An Empirical Ensemble Framework for Fake News Classification

*Travel*'s Practice at WSDM Cup 2019 Fake News Classification Challenge

Shuaipeng Liu, Shuo Liu, Lei Ren
Meituan-Dianping Group
{liushuaipeng, liushuo19, renlei04}@meituan.com

## ABSTRACT

With the rapid development of information technology, people now can get access to the overwhelming volume of online news content conveniently. However, fake news which provides inaccurate or misleading information, poses a big threat to the online news ecosystem and even the human civilization. In this paper, we proposed an ensemble framework to address the fake news classification challenge in ACM WSDM Cup 2019[1]. In our solution, we regarded this problem as the Natural Language Inference (NLI) task and proposed a novel empirical ensemble framework and finally our team *Travel* won the 2nd place with a weighted accuracy score of 0.88156 on the private leaderboard.

## CCS CONCEPTS

• **Computing methodologies →Classification and regression trees;**

## KEYWORDS

fake news classification, BERT, ensemble model

## 1 INTRODUCTION

ByteDance is a Chinese Internet technology company operating several machine learning-enabled content platforms. One of the challenges ByteDance faces is to combat different types of fake news, referring to all forms of false, inaccurate, or misleading information. As a result, ByteDance has created a large database of fake news articles, and any new article must go through a test for content truthfulness before being published, based on matching between the new article and the articles in the database. Articles identified as containing fake news are then withdrawn after human verification of their status. The accuracy and efficiency of the process, therefore, are crucial in regard to making the platform safe, reliable, and healthy. The company invites researchers and students in the community to take part in the following task[2]. Given the title of a fake news article A and the title of a coming news article B, participants are asked to classify B into one of the three categories: 1) agreed: B talks about the same fake news as A; 2) disagreed: B refutes the fake news in A; 3) unrelated: B is unrelated to A.

[1]http://www.wsdm-conference.org/2019/wsdm-cup-2019.php
[2] https://www.kaggle.com/c/fake-news-pair-classification-challenge/

This competition uses Weighted Accuracy as the evaluation metric which is described by the following function:

$$\text{WeightedAccuracy}(y, \hat{y}, \omega) = \frac{1}{n}\sum_{i=1}^{n}\frac{\omega_i(y_i = \hat{y}_i)}{\sum \omega_i}$$

Where y are the ground truth, $\hat{y}$ are the predicted results, and $\omega_i$ is the weight associated with the *i*th item in the dataset. The weights of the three categories, agreed, disagreed and unrelated are 1/15, 1/5, 1/16 respectively.

After analyzing the challenge task, we find that this task is similar to the Nature Language Inference (NLI) task in NLP, also known as recognizing textual entailment (RTE). The NLI task focuses on the problem whether a hypothesis can be inferred from a premise, requiring a deep understanding of the semantic similarity between the hypothesis and the premise [1-2]. There are many datasets having been built to study the NLI task, such as MNLI [3], QNLI [4], RTE [5], WNLI [6] and SNLI [7], and all of them are included in The General Language Understanding Evaluation (GLUE) benchmark dataset [8]. And to handle the NLI problem, various models have been proposed by researchers, including ESIM [9], OpenAI GPT [10] and BERT [11]. As far as we know, BERT, a pretrained language model, is the state-of-the-art model for the NLI task now.

In this paper, we introduce an empirical ensemble framework which has a three-level architecture to address the problem. The first level contains 25 BERT model with a blending ensemble strategy, the second level contains 6 traditional machine learning method with 5-fold stacking ensemble strategy, and the third level using a single Logistic Regression to generate the final result.

The rest of the paper is organized as follows: Section 2 introduces the dataset of the competition. In Section 3, we describe our solution which contains the model details. In Section 4, we show the experiment results of our model. Finally, we conclude our analysis of the challenge, as well as some additional discussions of the future directions in Section 5.

## 2 DATASET

For this challenge, the organizer has provided a training dataset and a testing dataset, in which the training dataset consists of 320,767 news pairs with 3 class labels (agreed, disagreed and

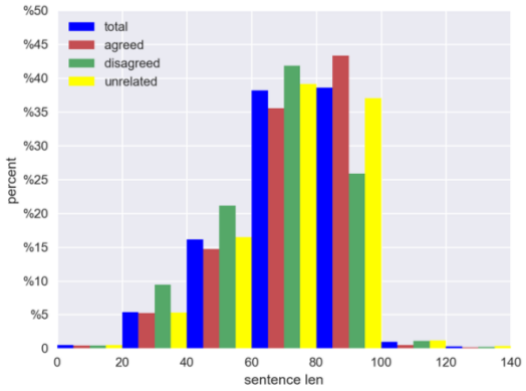


**Figure 1: Distribution of news title length in Chinese**

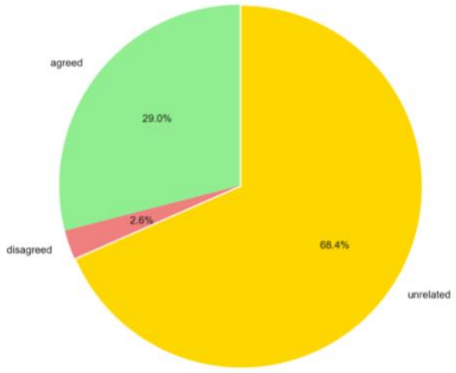


**Figure 2: Distribution of class labels in the training dataset**

unrelated), and the testing dataset contains 80,126 news pairs without labels. We are required to train models using the training dataset and predict the class label of the testing dataset. These news pairs are in both Chinese and English. Figure 1 shows the distribution of sentence length with different categories, and we can find that the length distribution of different categories is almost same. Figure 2 shows the distribution of category labels, and we can find that the distribution of class labels are fairly imbalanced, which increases the difficulty of the task.

## 3 OUR SOLUTION

This section describes the details of our solution including data augment, data preprocessing, base models construction and model ensemble. An overall framework and processing pipeline of our solution is showed in Figure 3. Our trained models and source code are publicly available on GitHub. [3]

### 3.1 Data Augment

Data augmentation is an effective way to alleviate the over-fitting problem by increasing the amount of training data, and can also alleviate the problem of data imbalance. In this challenge, we proposed a simple yet effective method for data augment, by using the semantic transitivity. As shown in Figure 4, if title A is agreed with title B, and title A is agreed with title C, then we can draw a conclusion that title B is agree with title C. In addition, if title A is agreed with title B, and title A is disagreed title D, then we can conclude that title B is disagreed with title D. We also generated new data by swap the sequence of the news pairs. In addition, exchanging the order of original news pairs simply can also generate more training data.

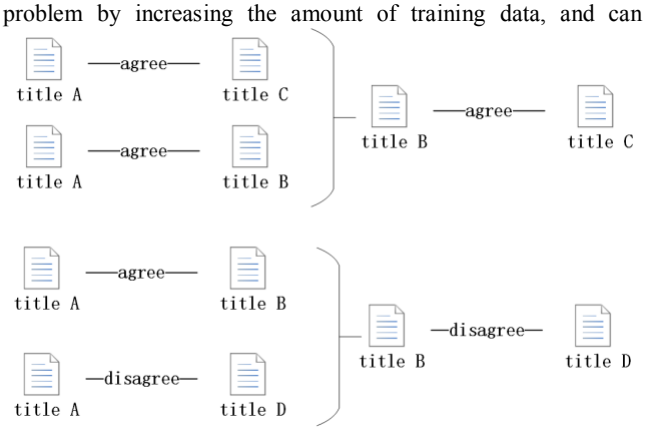


**Figure 4: Data augment strategy for generating news pairs**

### 3.2 Data Preprocess

#### Data preprocess

In the data preprocessing part, we first made a transformation between Traditional Chinese and Simple Chinese for the news titles, and then removed the stop words in the dataset. You can get the detail settings in our open source code.

### 3.3 Base Model

In the paper, we used BERT as the base model. BERT is a language representation model designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. Figure 5 shows the architecture of BERT for pre-training with a bidirectional Transformer.

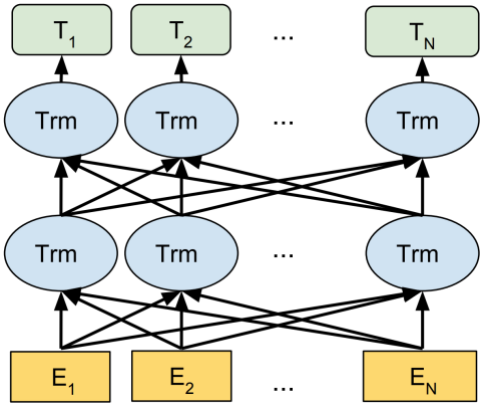


**Figure 5: Architecture of BERT for pre-training with a bidirectional Transformer**

[3] https://github.com/myeclipse/WSDM2019

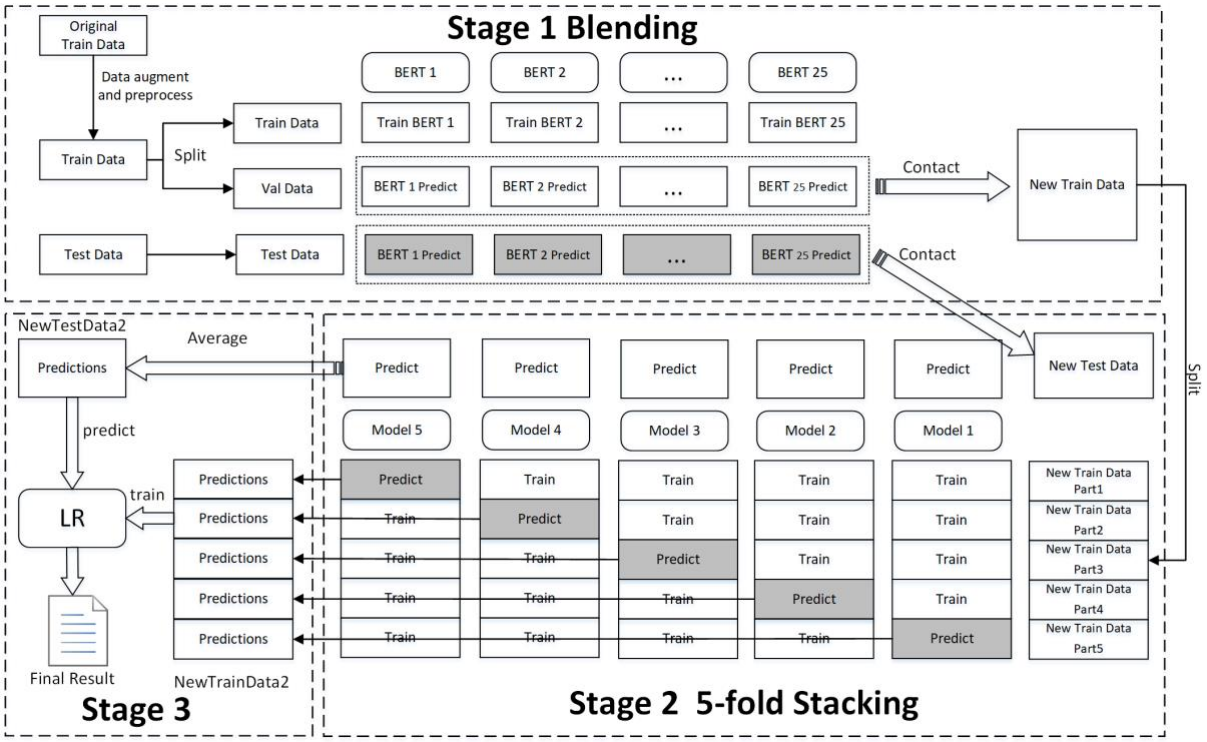


**Figure 3: An overall framework and pipeline of our solution for fake news classification**

In practice, we fine-tuned Google's pre-trained BERT models including both Chinese version and English version with just one additional output layer using different parameters, and then used it to classify the fake news. The final ensemble model combined 25 single BERT models with different configurations and you can get the detail settings in our open source code. The architecture of the single BERT model is shown in Figure 6.

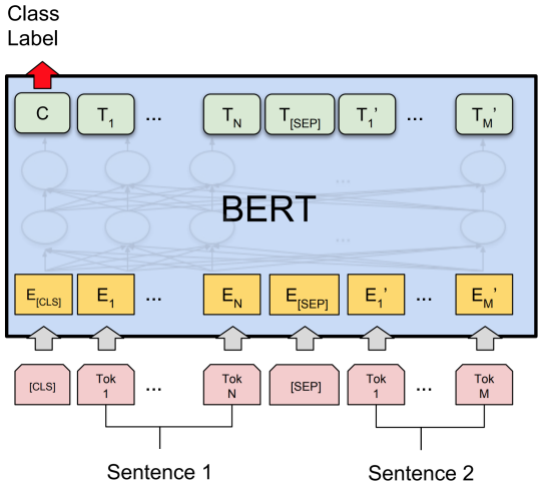


**Figure 6: Fake news classification model by incorporating BERT with one additional output layer**

## 3.4 Model Ensemble

As shown in Figure 7, we trained a three-level model to perform the fake news classification. In the first level, the data were fed into 25 BERTs, the outputs from the first level would be regarded as features for the second level to learn. Then in the second level, we trained six models including three SVMs, Logistic Regression (LR), K-Nearest Neighbor (KNN) and Naive Bayesian (NB), the outputs of those second level models would be the features for the third level model to learn, forming a stacking-based model ensemble. In the third level, a LR model was only used to generate the final result.

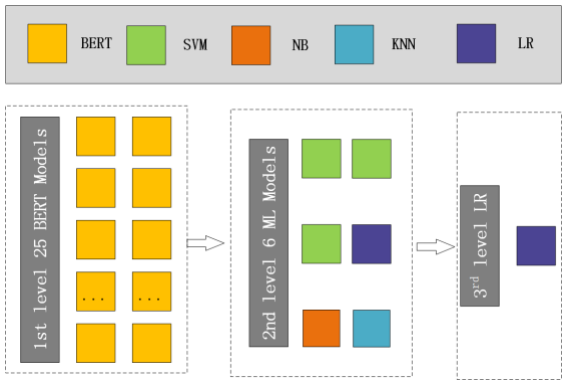


**Figure 7: Model Ensemble Architecture**

In our practice, the first level used the blending technique with a split of 95% training data to learn and 5% training data to predict, the reason of choosing blending rather than stacking is that the time-cost of BERT is very expensive. Then in the second level, we adopted the 5-fold stacking technique [12] to train and predict. In the third level, we used the 5-fold cross validation to avoid overfitting. The processing detail is illustrated in Figure 5.

## 4 EXPERIMENTS

Table 1 lists the results of various models described previously. On the private leaderboard, the best performance of the single base model among 25 BERTs can get a weighted accuracy of 0.86750, averaging of 25 BERTs can get 0.87700, weighted averaging of 25 BERTs can get 0.87702, and our empirical ensemble model gets the best performance of 0.88156. As shown in the table our empirical ensemble model can greatly improve the performance and our overall approach achieved 2rd place on the final leaderboard.

Table 1: Performance of Various Models

| Model | Weighted Acc on Private LB |
|---|---|
| Best Single base model | 0.86750 |
| Averaging of 25 BERT | 0.87700 |
| Weighted Averaging of 25 BERT | 0.87702 |
| Our Empirical Ensemble Model | 0.88156 |

## 5 CONCLUSIONS

In this paper, we have introduced an empirical ensemble framework for the Fake news Classification Challenge of the WSDM Cup 2019. Our team *Travel* was ranked the second place on the final leaderboard. In our solution, we first conducted data augmentation and data preprocessing, then we trained 25 BERTs as the base models. After that, we ensembled these base models using 5-fold stacking strategy to generate the probability of the test data. Finally, A LR model was trained using previous level's outputs, and then predicted the class labels of the testing data as the final result. While obtaining promising performance on the whole, our model still cannot handle some bad cases. We will leave these challenges for future work. For example, since Google's BERT model is pre-trained by Wikipedia data, to solve the data consistency problem, we can pre-train or continue pre-train the BERT model using news articles. What is more, we can design the new neural network architecture which can catch more complex signals of fake news.

## ACKNOWLEDGMENTS·

Firstly, we thank everyone associated with organizing and sponsoring the WSDM Cup 2019. Dataset was provided by ByteDance. Challenge was sponsored and managed by the 12th ACM International Conference on Web Search and Data Mining (WSDM 2019). Competition platform was hosted by Kaggle. Then, we are very grateful to Zhongyuan Wang, Huixing Jiang and Fuzheng Zhang for their great support during the challenge.

## REFERENCES

[1] Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge, Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment. Springer, Berlin, Heidelberg, 177-190.

[2] Bowman S R, Angeli G, Potts C, et al. 2015. A large annotated corpus for learning natural language inference. In proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[3] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.

[4] Rajpurkar P, Zhang J, Lopyrev K, et al. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250,

[5] Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In TAC. NIST.

[6] Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In Aaai spring symposium: Logical formalizations of commonsense reasoning, volume 46, page 47.

[7] Bowman, Samuel R., et al. 2015. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326.

[8] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461.

[9] Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. 2016. Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[12] David H. Wolpert. 1992. Stacked generalization. Neural Networks (1992). https: //doi.org/10.1016/S0893-6080(05)80023- 1