Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News

The first-place entry for Fake News Classification at WSDM Cup 2019

Lam Pham
DataLabs, VNG Corp.
HCMC, Vietnam
lampt@vng.com.vn

ABSTRACT

As fake news is sowing confusion and posing a big threat to our human being society, the accurate and efficient solutions of fake news detection become crucial in order to make the global content platform like at ByteDance safe, reliable, and healthy. In this paper, we describe our winning approach for identifying fake news. In particular, we will explore the application of Bidirectional Encoder Representations from Transformers (BERT) [2] for encoding a news title pair, incorporating and transforming it into a new representation space for building and fusing downstream classifiers such as gradient boosting trees and neural networks. This proposed solution achieved weighted accuracy score of 88.29 in the private leader board, and was selected as first place submission, namely IM (Incredible Machine) team. [1]

KEYWORDS

Fake News Classification, Content Platform, WSDM Cup

ACM Reference format:
Lam Pham. 2019. Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News. In *Proceedings of WSDM conference, AU, February 2019 (WSDM'19),* 4 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the development of the Internet Technology, the popularity of mobile devices and the increasing amount of information generated on social media platforms, people can easily consume and share various content in different digital forms such as texts, voices, videos. Despite the advantages provided by social media, the quality of news on social media is not as good as traditional news organizations [6]. However, because it is much faster and easier to disseminate through social media, large volumes of fake news, i.e., those news articles with intentionally false, inaccurate and misleading information, are produced online for a variety of purposes, such as financial and political gain. The extensive spread of fake news can have serious negative effects on individuals and organizations. To mitigate this, scientists at ByteDance have created a large-scale database to store existing fake news articles. Any new article must go through a test on the truthfulness of content before being published. Concretely they conduct text matching between the checking article and the articles in the database. Articles identified as containing fake news will be withdrawn after a human verification.

In Fake News Classification Challenge[2], which is created by ByteDance, participants were challenged to build an algorithm that predicts whether a news title pair is agreed, disagreed or unrelated. To address this challenge, we introduce a novel ensemble approach using diversified learners such as gradient boosting trees and neural nets with transferring BERT encoding of news title pair input (in Chinese), transforming and combining it with some handcrafted features including lengths, pairwise distances from a bag of character representation, heuristic statistics. Our approach achieved 88.298 and 88.098 weighted accuracy in the private and public leader boards, respectively.

2 DATASET

We have utilised the training data which contains 320, 767 news title pairs in Chinese and translated-by-machine in English. However, we only leverage the Chinese inputs which are in high quality as recommended in this competition guidance. The size of test set is 80, 126 entries which are used to evaluate our approach on public and private leader boards with 25% and 75% of total samples, respectively. Given the news title pair A and B, where A is title of fake news, the task is classifying news title B into three folds:

- agreed: B talks about the same fake news as A
- disagreed: B refutes the fake news in A
- unrelated: B is unrelated to A

2.1 Exploratory Data Analysis

We first do exploratory analysis on given training and testing data and figure out that there are some misleading labels for the same title pairs in the Table 1

Secondly, we have observed the imbalance of three categories in the training data. In particular, unrelated pairs account for approximately 70% of the dataset. Additionally, we recognize that data is well shuffled for all classes. It means that there is no leakage in data preparation.

Finally, it is useful to swap the news title pair order from (A B) to (B A) to train a set of twin models not only for enriching the

[1]The source code is available at https://github.com/lampts/wsdm19cup

WSDM'19, February 2019, AU

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[2]https://www.kaggle.com/c/fake-news-pair-classification-challenge

Table 1: Noisy labels for same news pair inputs

tid1	tid2	title_zh	translated	label
5647	5647	"博鳌亚洲论坛"辟谣：与"博鳌亚洲区块链...	"Boao Forum for Asia": It has nothing to do wi...	unrelated
9990	9990	2020年取消农村低保补助？专家：补助只会扩大...	In 2020, the rural subsidies will be abolished...	disagreed
26104	26104	"小黄车传艾滋病"？谣言！都是谣言！	"Yellow Car AIDS"? Rumors! Rumors!	unrelated

data inputs but also for increasing the variance of doing ensemble in final prediction.

2.2 Evaluation Metric

In order to reduce the bias of model performance, Weighted Categorization Accuracy is used to evaluate the model performance. The weights of agreed, disagreed and unrelated categories are $\frac{1}{15}$, $\frac{1}{5}$, $\frac{1}{16}$, respectively.

Precisely, Weighted Categorization Accuracy can be generally defined as:

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

where y are ground truths, $\hat{y}$ are the predicted results, and ω_i is the weight associated with the ith item in the dataset.

3 PROPOSED APPROACH

In this section we elaborate our proposed solution including feature engineering and modelling.

3.1 BERT encoding of news pair

Sentence encoding or embedding is an upstream task required in many NLP applications, e.g. sentiment analysis, text classification. The goal is to represent a variable length sentence into a fixed length vector. For example the sentence "have a good day" can be encoded as $[0.2, 0.3, 0.7]$. Each element of the vector should represent some semantics of the original sentence. Recently, Bidirectional Encoder Representations from Transformers (BERT) [2] which achieves a big milestone in NLP field is pre-training language representations developed by Google. It consumes an enormous amount of plain text data publicly available on the web and is trained in an unsupervised manner.

To encode a news title pair into 768 dimensional semantic space, we leverage the base 12 layer version for a language model in Chinese. We use the layer before the final hidden layer to preserve much more information for our fine tuned models later. To see that more clearly, Figure 1 is a visual representation of $15K$ samples, where we randomly select 5000 pairs for each category, then get a vector from a layer before the final hidden layer, and finally reduce it into two dimensions via PCA.

3.2 Handcrafted Feature Engineering

Before extracting features, we have done very simple pre-processing for Chinese such as lowering, adding space for punctuation. After that features such as statistic, text based, graph based and KNN based are extracted from pre-processed data.

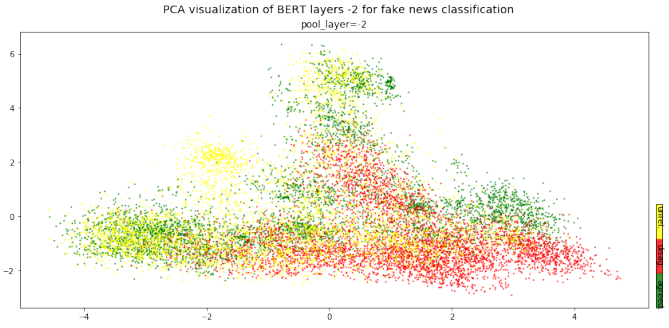

Figure 1: 15K pairs of 3 fake news categories via PCA

3.2.1 Statistics. We first have 15 statistic features for string pair matching based on lengths, common set of words, stop words, tokens, characters. This feature set contributes some predictive features in the top of most important features selected by tree based models. As shown in Table 2, the overlap ratio of string matching is second rank feature.

3.2.2 Text based. For textual inputs, we apply traditional NLP techniques such ngrams of characters or words to vectorize text into numerical space. The news title pair then is transformed using pairwise distances such as cosine, euclidean, city-block, jaccard or just simply using summation or subtraction.

3.2.3 Graph based. Interestingly, there is creative representation of text inputs as graph network. Here every news title is a node of graph and a news pair indicates an edge between the two titles. We can use this graph network to enrich our feature pool including count and ratio of intersections of neighbours, minimum and maximum frequency of nodes, minimum and maximum of kcore for given pair of nodes.

3.2.4 KNN features. After having a 768-dimensional vector for each news title pair from BERT embedding, we first apply dimension reduction via SVD with 168 dimensions then build another feature extractor via KNN classifier. Based on this model, we extract very powerful features including object fraction of 32 nearest neighbours in each category, the longest streak of same label within 32 nearest neighbours and minimum or mean distance to object of each label group. Those features are in fact our top features for tree based models as given in Table 2.

3.3 Modelling

As shown in Table 1, there are some misleading labels in the training data where the same titles have different labels. To solve this issue, we propose a bagging method which not only leverages all training data but also reduces noises in the training data. We can

Table 2: Top features performance by tree based models

feature	description	importance by split
knn_16	minimum normalized distance to object of unrelated category	1592
stat_13	overlap ratio of string matching	1214
stat_14	partial overlap ratio of string matching	1006
stat_11	token set ratio matching	888
knn_14	minimum normalized distance to object of agreed category	869
stat_15	longest sub-string ratio matching	796
knn_9	fraction of object in unrelated category of 32 nearest neighbours	793
knn_25	mean distance to 3 nearest neighbours of unrelated category	763
knn_13	minimum distance to object of unrelated category	746
谣	word Rumours	743

see the overall solution architecture which consists of feature pool generation, base model pool and blending in Figure 2. We use several base models including Logistic Regression, Gradient Boosting Trees and Neural Networks in our model pool for combining the final score later.

3.3.1 Logistic Regression. Our first baseline model is Logistic Regression using Scikit-Learn package [5]. This fast trained model achieved weighted accuracy 82.3 using stratified 5-fold cross validation.

3.3.2 Tree-based Models. Gradient boosting is a powerful meta-algorithm used to reduce prediction bias. Recently, Catboost [3] and later LightGBM [4] have gained increased popularity and attention due to their advantages of fast processing speed and high prediction performance. We utilize Catboost and LightGBM to build 9 models with different feature sets. In comparison to our baseline model their performance are relatively high (about 86.2) and stable on public and private leader boards. Moreover, these tree based models are really helpful in contribution of final ensemble as they diversify and improve our score performance. Therefore, we keep them in the final blending stage.

3.3.3 Neural Networks Models. Undoubtedly deep learning (DL) added a huge boost to many AI applications such as image recognition, speech recognition, and text understanding [1]. For this competition, we fine-tune neural networks which have one dense layer in order to learn the interaction between the BERT encoding and handcrafted features described above. We also swap the input pair to train different models for maximizing the variance and leveraging more data samples. Our neural networks outperform other tree based models and give a big advantage for final prediction. As it points out in the Table 3, the average prediction of all neural networks gives us very competitive scores which are 87.84 and 87.91 on private and public leader boards, respectively.

The hyperparameter values we have to train the neural networks are as follows:

- Batch size: 16, 32
- Learning rate(Adam): 2e-5
- Number of epochs: 3,4,5
- Maximum sequence lengths: 156, 168, 176, 182

We think that the training size is big enough so we do not need to pay more attention on hyperparameter tuning.

Table 3: Top model performance

algorithm	private score	public score
Best single tree-based	85.77%	85.96%
Best single nnet	87.21%	87.18%
Avg best of tree-based and nnets	87.73%	87.92%
Avg 9 tree based models	86.17%	86.38%
Avg 18 nnet models	87.84%	87.91%
Avg 18 nnets + 9 trees + 1 logistics	88.28%	88.19%

3.3.4 Final Solution. Our final solution is the blended predictions of 18 neural networks, 9 tree based models and logistic regression. All neural networks models can be trained simultaneously using multiple GPU devices. For training tree based models, we use multi-core CPU based version using different cross validation strategy including random seed number, number of folds. We then simply average all predictions to have the final prediction.

4 LESSONS LEARNED

We found that the distribution of target variable is different in train, public and private sets. This discrepancy can be achieved using probing leader-board scores. As shown in Figure 3, the unrelated class is reduced significantly in private set. It means that simple blending approach probably make our model more robust and reliable to the distribution shift.

Moreover, we have extracted the second last layer from pre-trained BERT to get a representation of news pair. This can be a key for this solution to get top scores. In the future, we can have more investigations on using more layers or combining some layers to see whether we can gain more performance boost.

5 CONCLUSION

In this paper, we presented our winning solution for fake news classification. The proposed approach which uses compelling machine learning algorithms such as light gradient boosting (LGB), Catboost (CB), Neural Networks (NNets) with novel features in text mining such as sentence pair encoding and transformation gives outstanding performance. We hope that our approach will probably provide a empirical benchmark for identifying fake news as well as other text matching problems in natural language understanding for news media platforms.

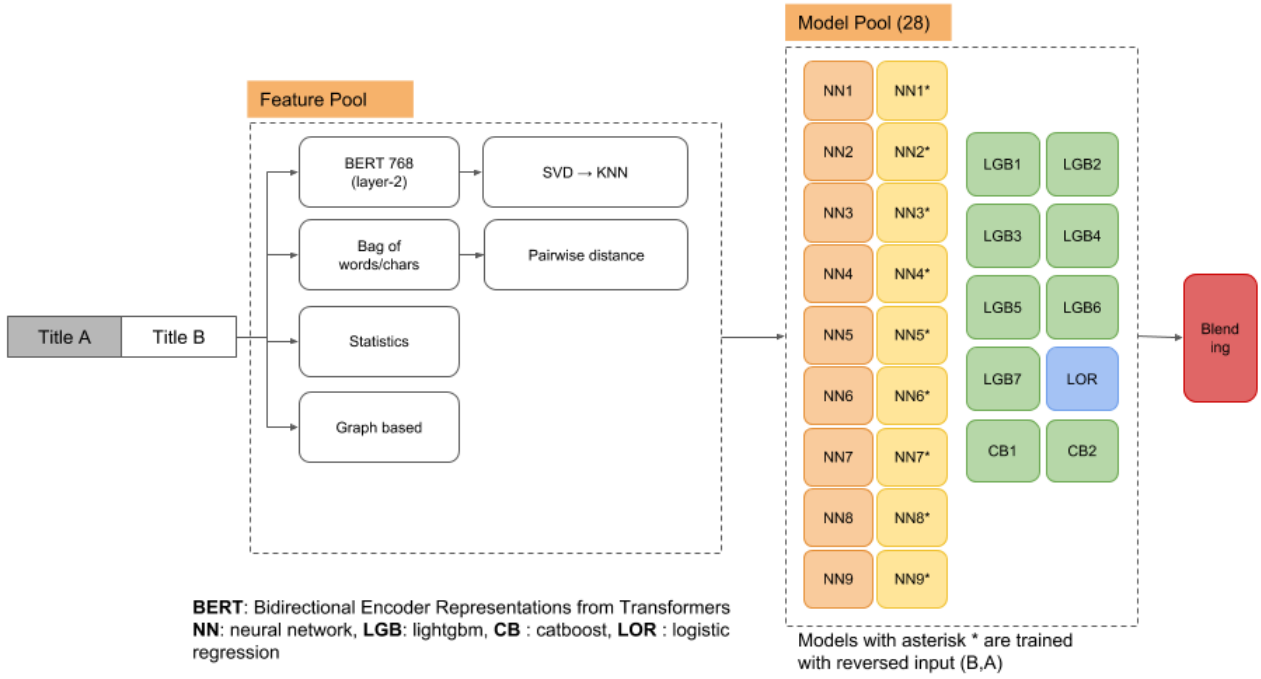

Figure 2: Fake News Classification Modelling

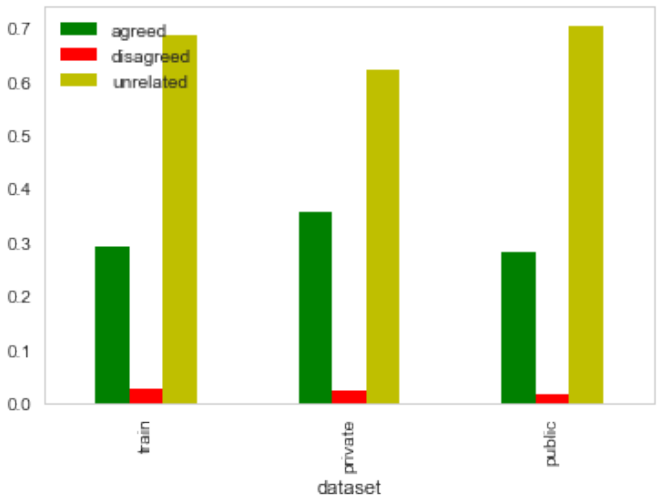

Figure 3: Class distributions in different data sets

REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (Aug. 2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev. 2017. Fighting biases with dynamic boosting. *CoRR* abs/1706.09516 (2017). arXiv:1706.09516 http://arxiv.org/abs/1706.09516

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3149–3157. http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[6] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *CoRR* abs/1708.01967 (2017). arXiv:1708.01967 http://arxiv.org/abs/1708.01967