HAHA at FakeDeS 2021: A Fake News Detection Method Based on TF-IDF and Ensemble Machine Learning

Kun Li

School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China
2106967047@qq.com

Abstract. This paper describes our participation in the FakeDeS [5] Task at Iberlef 2021: Fake News Detection in Spanish. Base on this task, we propose the classic TF-IDF feature extraction technology and Stacking ensemble learning method base on weak classifiers. It not only analyzes the content of the news, but also combines effective information such as publishers and topics to improve the performance of our model. We used five machine learning models, and achieved very competitive results on both the validation set and the test set, and got the second place in the final evaluation phase.

Keywords: Fake News Detection, TF-IDF, Machine Learning, Ensemble Model

1 Introduction

Fake news refers to a kind of public opinion that uses false information to deceive the parties in order to achieve a certain purpose. It fails to truly reflect the original appearance of objective things and contains false elements. The information provided by fake news is designed to manipulate people for different purposes: terrorism, political election, advertising, satire, etc. In social networks, fake news spreads in seconds among thousands of people and research has shown that misinformation spreads faster, farther, deeper, and more widely than true information [12], so it is necessary to develop tools to help control the amount of fake news on the network.

A few years ago, the method of detecting fake news was mainly to analyze the effective features from various sources, including the content of the text, user data and the form of news dissemination. It mainly distinguishes true and false news from the aspect of language features, such as writing style and special headlines [7], vocabulary and syntactic analysis [10]. In addition to language features [3], some studies have proposed classification schemes on user features

IberLEF 2021, September 2021, Málaga, Spain.

and time features [7]. Recent fake news detection methods mainly use machine learning and deep learning techniques, with special attention to language-based methods [8, 11, 14, 15]. Some people use TF-IDF feature extraction technology for fake news detection and have achieved good results [1, 13, 2].

A fake news detection system is designed to help user detect and filter potentially deceptive news. A predictive method of deliberately misleading news is based on the analysis of the real and deceptive previously censored news, that is, the annotated corpus. Text is main carrier of news information, and the study of news text helps to effectively identify fake news. The specific task is: given the text of a news event, determine whether the event is real news of fake news. For the evaluation of systems, we will use a new testing corpus containing news related to COVID-19 and news from other Ibero-American countries. Its availability will introduce two main challenges to the task: thematic and language variation. Our systems need to take into consideration that part of the testing corpus contains news in a topic that does not exist in training corpus, likewise, we should take into account that the other part of the testing corpus contains news in a different variation of the Spanish that is in training corpus. This paper proposes a method for fake news detection: A fake news detection method based on TF-IDF and ensemble machine learning.

TF-IDF has the characteristics of simplicity, fast calculation. and it performs well for processing long texts. The Section 2 introduces the corpus and analyzes the composition and distribution of the data. The third section introduce the methodology, data processing methods, feature extraction methods, the base model used, and the final ensemble model. The experiments sett and results are presented in Section 4. Finally, Section 5 outlines the final conclusions and future work.

2 Corpus Description

The Spanish fake news corpus [9] is news collected from several online sources: existing newspaper websites, media company websites, special websites dedicated to verifying fake news, and websites designated by different reporters as regular fake news publications. All these articles are written in Mexican Spanish. The corpus collected 971 news items from different sources from January to July 2018: 971 news items were divided into training set and test set. Among them, there are 676 pieces of data in training set and 295 pieces of data in the test set. Only two categories (True of Fake) are considered for the marking of the corpus, and the specific conditions of each piece of data are as follows:

- Category: Fake/ True.
- Topic: Science/ Sport/ Economy/ Education/ Entertainment/ Politics, Health/ Security/ Society.
- Headline: The title of the news.
- Text: The complete text of the news.
- Link: The URL where the news was published.

Among them, the number of fake news and real news is fairly balanced. And the number of fake news and real news in the Topic column is almost the same. However, there is a big gap between the authenticity of news published on different websites. Some websites are almost all fake news, and there are also websites that are all real news. This provides a good idea for our feature extraction. We will consider the impact of "Link" and "Topic" on the classification results when we do experiments.

3 Method and Technology

This section includes 4 parts: data preprocessing, feature extraction methods, classification models, and ensemble model methods.

3.1 Data Preprocessing

In order to get better results, data preprocessing is essential. And data preprocessing is usually the first step in natural language processing tasks. First intercept the most critical information in Link, and the result after interception looks like this: www.elruinaversal.com. Because we have analyzed the sources of fake news and found that almost all the news on some websites is fake news. Then we observe the data and find that each row consists of Category, Topic, Source, Headline, Text and Link. All the contents in Category, Topic, Source, Headline, Text and Link are merged as our new input. Finally, data cleaning is performed on the merged input data. Perform data cleaning on the merged data, use regular expressions to remove links, special symbols, punctuation, etc. According to the length of the text, the length of the longest text is 2578 and the shortest text is 31. So we decided to use nltk to remove the stop words in the text. The longest text length after removing the stop words is 1379, and the shortest is 18. Removing stop words in the text will reduce the effect of the model, and not removing stop words will improve the performance of the model. However, we have proved through experiments that removing stop words will reduce the performance of the model, which will be explained in subsequent experiments. And we will verify it in the next experiment. The data processing is: 1) merge all columns + remove stop words: 2) merge all columns + without remove stop words: 3) only Text + remove stop words: 4) only Text + without remove stop words.

3.2 Feature Extraction

The method base on news content focuses on extracting various features of fake news content, including knowledge-base and style-based features. This paper mainly uses two methods to extract text features: 1) LabelEncoder; 2) TF-IDF.

- LabelEncoder: We use Sklearn's LabelEncoder method to hard-code text features, that is, to encode discrete numbers or text, and convert the discrete data to numbers between (0, n-1), where n represents different data

values. We performed LabelEncoder on the Topic and Source features. The LabelEncoder method also played a very good role in the experiment.

- TF-IDF: Term Frequency-inverse Document Frequency (TF-IDF) is a statistical analysis method for keywords, used to evaluate the importance of a word to a document set or a corpus. The importance of a word is positively correlated with the number of times it appears in the article, and negatively correlated with the number of times it appears in the corpus. TF-IDF can effectively avoid the influence of commonly used words on keywords and improve the relevance between keywords and articles. TF refers to the total number of times a word appears in the article. This indicator is usually normalized and defined as the number of times a word appears in the article divided by the total number of words in the article, which can prevent the result from being biased towards too long document (The same word usually has a higher word frequency in a long document than in a short document). IDF refers to the frequency of reverse documents. The fewer documents that contain a word, the greater of the IDF value, indicating that the word has a strong ability to distinguish. Using TF-IDF can well extract text features in Spanish news. For texts with a length of several thousand, TF-IDF is better than RNN and other neural network model in extracting features of long texts. And for the challenge of changing language, TF-IDF can easily solve it.

3.3 Base Classification Model

We use five basic weak classifiers as our base model.

- LogisticRegression (LR): Logistic regression is used to discover the connection between features and output results, used for classification problems in supervised machine learning algorithms, and has a close relationship with neural networks. Neural networks can be regarded as multiple logistic regression classifiers Stacked. Logistic regression can be used for binary classification problems and multi-classification problems.
- SGDClassifier (SGDC): Mainly used in large-scale sparse data problems. It is a collection of linear classifiers trained with stochastic gradient descent algorithm, It is a linear (soft interval) support vector machine classifier by default, which is logistic regression in this article.
- PassiveAggressiveClassifier (PAC): It is a classic online linear classifier, which can continuously integrate new samples to adjust the classification model and enhance the classification ability of the model. It can perform feature extraction on streaming data, and can perform incremental learning.
- RidgeClassifier (RC): This classifier uses penalized least square function to adapt to the classification model. The loss function used by RidgeClassifier can make different calculation performance profiles.
- LnearSVC (LSVC): A linear classification supports vector machine is implemented by liblinear, which can be used for two-class classification or multi-class classification.

3.4 Ensemble Model

LightGBM [6]: The model is a fast, distributed, high-performance gradient boosting framework base on decision tree algorithm. It can be used in sorting, classification, regression, and many other machine learning tasks. LightGBM is an improvement of the GBDT algorithm [4]. LightGBM didn't use the traditional pre-sorting idea, but optimizes the histogram of the eigenvalues. The weak classifier is used to iteratively train to obtain the optimal model, which has the advantages of good training effect and not easy to overfit. The train method is GBDT: it is an algorithm that classifies or regresses data by using an additive model and continuously reducing the residuals generated during the training process.

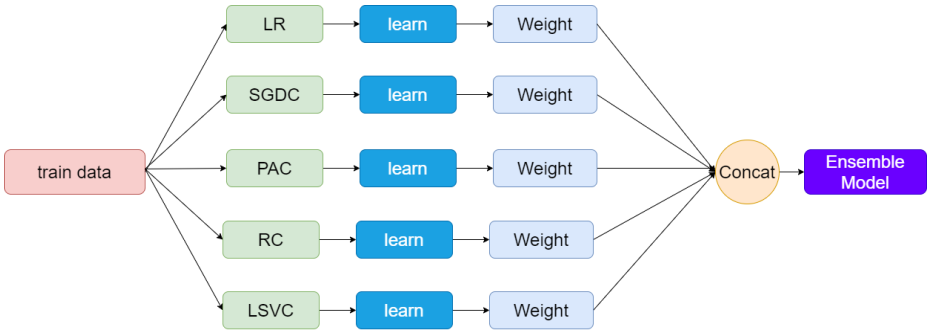

Fig. 1. Combine the processed data of the five basic models and then perform ensemble learning.

Choose the above LogisticRegression (LR), SGDClassifier (SGDC), PassiveAggressiveClassifier (PAC), RidgeClassifier (RC), LnearSVC (LSVC) as the weak base model. Figure 1 shows the use of the Stacking method in the ensemble learning method to predict all the trained base models on the entire training set, and each base model will get a classification prediction result. For each base model, we train out model by using 5-fold cross-validation, concatenate the classification prediction results after each base model training, and finally send all the features to the final LightGBM model for training.

4 Experiments

4.1 Experimental Setup

First, data processing is performed, and then all the hyperparameter of the experiment are introduced. Almost all base model use default parameters, and each base model uses 5-fold cross-validation for training; the LightGBM model also uses 5-fold cross-validation for training. The training model of the LightGBM

model is GBDT; the learning rate is set to 0.01; the maximum number of iterations num-boost-round is set to 10000; the progress is displayed every 50 iterations. Finally, the model outputs the final accuracy rate and and F1-macro. The hyperparameters of each classifier are as follows: LR (random-state=1017, C=3), SGDC (random-state=1017, loss='log'), PAC (random-state=1017, C=2), RC (random-state=1017), LSVC (random-state=1017).

4.2 Result

We will evaluate the model from F1 measure and accuracy on the "fake" class. The results of the base model and ensemble model on the training data shown in Table 1, where "Merge-All" is to concatenate all the information, and "Stopwords" is removing stop words from the text. Accuracy and F1-macro are both results obtained on the validation set. From Table 1, we can see that the performance of the ensemble model is always better than that of the base model,

Table 1. Results for the development data set of fake news by using different classifier models, and preprocessing techniques.

Model	Merge All	Stopwords	Accuracy	F1-macro
LR	Yes	Yes	0.876	0.880
SGDC	Yes	Yes	0.883	0.890
PAC	Yes	Yes	0.885	0.892
RC	Yes	Yes	0.873	0.878
LS	Yes	Yes	0.879	0.883
Ensemble	Yes	Yes	**0.911**	**0.913**
LR	Yes	No	0.859	0.861
SGDC	Yes	No	0.865	0.871
PAC	Yes	No	0.870	0.875
RC	Yes	No	0.862	0.865
LS	Yes	No	0.865	0.869
Ensemble	Yes	No	**0.902**	**0.904**
LR	No	Yes	0.808	0.805
SGDC	No	Yes	0.811	0.816
PAC	No	Yes	0.805	0.813
RC	No	Yes	0.817	0.815
LS	No	Yes	0.812	0.812
Ensemble	No	Yes	**0.882**	**0.884**
LR	No	No	0.822	0.819
SGDC	No	No	0.822	0.828
PAC	No	No	0.827	0.832
RC	No	No	0.819	0.816
LS	No	No	0.828	0.828
Ensemble	No	No	**0.898**	**0.900**

regardless of whether the merging and stop words are removed. For the same model, the highest accuracy can be obtained by performing two data processing methods at the same time. 91.1% and the highest F1-macro score 91.3%. At the same time, removing the stop words actually weakens the performance of the model, and merging all column information can significantly improve the performance of the base model, and it also improves the ensemble model to a certain extent.

Table 2. The results of our model on the official test sets.

Model	F1-macro	rank
Ensemble	0.7548	2

We can get that in the same base model, merge can play a certain role in improving the accuracy of the model and the F1-macro score. Without the merge, the accuracy of the model is significantly reduced, and the remove stop words from the text will reduce the performance of the model. At the same time, no matter how the data is processed, the ensemble model is better than the base model in accuracy, and the F1-macro score of the ensemble model performs better with merged. Among them, the accuracy of the ensemble model is at least 2.6% higher than that of the base model, and at most 7%. The F1-macro score is improved by at least 2.1% and at most 6.8%. This fully illustrates that our ensemble learning model plays a very good role in improving the performance of the model. The more information the model inputs, the better the performance of the model, but too much data will affect the efficiency of model operation.

On the test data set, we only got two results: 1) not merging and removing stop words, 2) merging and not removing stop words. The F1-macro score of the first type is only 0.6975, and the F1-macro score of the second type reaches 0.7548 show in Table 2, which once again shows that data processing and ensemble learning methods can effectively improve the performance of the model.

5 Conclusions and Further Work

This article describes the fake news detection classification task in the IberLEF 2021 task. We used some classic feature extraction methods and machine learning techniques, and achieved high performance on the development set through ensemble methods. Compared with other deep learning and machine learning methods, the performance on the test set is also very competitive. Compared with MEX-A3T 2020 [13], the accuracy rate on the verification set has increased by about 8%, and the F1-macro score has increased by 6%. Compared with last year's best papers, the results of our model are also very competitive. The best F1-macro score we got on the test set was 0.7548, which was the second place. Due to changes in language and tweet content, the performance on the development set is still lower

The future work is to explore more advanced technologies, use better feature extraction methods, and achieve better results in the next competition. Secondly, we also plan to apply our model to other languages and better solve the current flood of Covid-19 information. Finally, it is harmful to treat all news from a link as fake news or real news. We will solve this problem in future work.

Acknowledgments

We would like to thank the organizers for the opportunity and organization of this task, as well as teachers and seniors for their help. Finally, we would like to thank the school for supporting my research and the patient work of future reviewers.

References

1. Ahmed, H., Traore, I., Saad, S.: Detecting opinion spams and fake news using text classification. Security and Privacy **1**(1), e9 (2018). https://doi.org/https://doi.org/10.1002/spy2.9, https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9
2. Arce-Cardenas, S., Fajardo-Delgado, D., Carmona, M.Á.Á.: Tecnm at MEX-A3T 2020: Fake news and aggressiveness analysis in mexican spanish. In: IberLEF@SEPLN. CEUR Workshop Proceedings, vol. 2664, pp. 265–272. CEUR-WS.org (2020)
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web. p. 675–684. WWW '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1963405.1963500, https://doi.org/10.1145/1963405.1963500
4. Friedman, J.H.: Greedy function approximation: A gradient boostingmachine. The Annals of Statistics **29**(5), 1189 – 1232 (2001). https://doi.org/10.1214/aos/1013203451, https://doi.org/10.1214/aos/1013203451
5. Gómez-Adorno, H., Posadas-Durán, J.P., Bel-Enguix, G., Porto, C.: Overview of fakedes task at iberlef 2020: Fake news detection in spanish. Procesamiento del Lenguaje Natural **67**(0) (2021)
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 3149–3157. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
7. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining. pp. 1103–1108 (2013). https://doi.org/10.1109/ICDM.2013.61
8. Oshikawa, R., Qian, J., Wang, W.Y.: A survey on natural language processing for fake news detection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6086–6093. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.747

9. Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G., Escobar, J.J.M.: Detection of fake news in a new corpus for the spanish language. Journal of Intelligent & Fuzzy Systems **36**(5), 4869–4876 (2019)
10. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 231–240. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1022, https://www.aclweb.org/anthology/P18-1022
11. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl. **19**(1), 22–36 (Sep 2017). https://doi.org/10.1145/3137597.3137600, https://doi.org/10.1145/3137597.3137600
12. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). https://doi.org/10.1126/science.aap9559, https://science.sciencemag.org/content/359/6380/1146
13. Zaizar-Gutiérrez, D., Fajardo-Delgado, D., Carmona, M.Á.Á.: Itcg's participation at MEX-A3T 2020: Aggressive identification and fake news detection based on textual features for mexican spanish. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) colocated with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 258–264. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2664/mexa3t_paper4.pdf
14. Zhang, X., Ghorbani, A.A.: An overview of online fake news: Characterization, detection, and discussion. Information Processing Management **57**(2), 102025 (2020). https://doi.org/https://doi.org/10.1016/j.ipm.2019.03.004, https://www.sciencedirect.com/science/article/pii/S0306457318306794
15. Zhou, X., Zafarani, R., Shu, K., Liu, H.: Fake news: Fundamental theories, detection strategies and challenges. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. p. 836–837. WSDM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/836–837, https://doi.org/10.1145/3289600.3291382