

Greeny at Factify 2022: Ensemble Model with Optimized RoBERTa for Multi-Modal Fact Verification

Wei Bai¹

¹University of Electronic Science and Technology of China, Chengdu, China

Abstract

In recent years, social media is becoming the main channel for people to obtain news, but also promotes the spread of fake news. Under the trend of rich media of social media, fake news gradually change from single text to multi-modal form, so multi-modal fake news detection is receiving more and more attention. First, we combine the pre-trained model Robustly Optimized Bert Pretraining Approach (RoBERTa) and other methods such as Bi-directional LSTM (BiLSTM) and UER by using the text and OCR information. The above-mentioned methods are trained as part of our ensemble model, together with semi-supervised training, are weighted to generate our final results. In the multi-modal model, we use RoBERTa and ResNet to extract text and image features respectively, and use Light Gradient Boosting Machine (LightGBM) to classify them. Finally, we fuse text-based and multimodal-based results and take the best-performing one. In the competition, our weighted average F1 score has reached 0.7428, achieving 6th place in FACTIFY.

Keywords

Fake news, Fact Verification, Multimodality, RoBERTa model

1. Introduction

Fake news, as a kind of false information deliberately created for political or economic purposes, has the characteristics of content hunting and fast dissemination. The proliferation of fake news not only triggers a storm of public opinion, but also can manipulate public events, which has a more direct harm to society than rumors [1]. The emergence of social media has greatly reduced the cost of spreading fake news. The widespread use of social media platforms represented by microblogs and Twitter has facilitated the fabrication and distortion of objective facts by manipulators of public events. Meanwhile, the social networks encourage users to produce their own content and publish, share, communicate and spread it through online platforms, making it more difficult to control fake news [2]. In 2020, a global networking overview by 'We Are Social' reports that the number of social media users in the world has reached 3.8 billion, nearly half of the world's population. Studies show that fake news spreads faster and more widely compared to real information [3]. During the 2016 U.S. presidential election, a large number of fake news spread widely on Facebook and Twitter, and were even alleged to have seriously influenced the outcome of the U.S. election [4]. It is undeniable that the ties between social media and news


De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

✉ cellurbw@gmail.com (W. Bai)

ORCID 0000-0002-4456-4532 (W. Bai)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

are getting more and more complicated and close. How to prevent social media from becoming a breeding ground for fake news, especially to stop the series of collateral damage caused by the manipulation of public events through social media platforms, has become a social issue worth exploring during the current global outbreak of COVID-19.

Although local governments now are paying more attention to the detection of fake information, the scale of data and the huge number of users hinder experts from correcting the inaccurate information and fake content in the information. It is difficult for the public to determine the authenticity of the information after obtaining the information. Many users judge the authenticity of the information based on their own understanding and cognition, rather than its source. A global study published by Edelman found that news and information search engines (63%) are trusted more than traditional media such as newspapers and television (58%) [5].

Nowadays, the rapid development of artificial intelligence technology brings hope for automatic detection of fake news. And the deep integration of natural language processing, computer vision and deep learning technology is developed to broaden the path of fake news detection. Multi-modal machine learning, as one of the advantageous methods used by deep learning techniques for feature representation, can take advantage of the complementarity between multiple modalities and circumvent the redundancy between modalities to achieve deep revelation and description of fake news features. Nowadays, limited by the text length of social media platforms, users often use multimedia forms of pictures and videos to enrich news stories and increase the expressive power of content to attract wider attention and dissemination [6]. Due to the heterogeneity of fake news content, multi-modal machine learning can be used to extract the relationship between modalities and perform deeper feature extraction and classification of text and image information in fake news [7].

In the rest of this paper, we organize the content as follows. Related work of fake news verification will be presented in Section 2. Section 3 introduces data description and the methodology of our models. Experimental results are discussed in Section 4. We also present the conclusion of our work at the end of paper.

2. Related works

Early researches on fake news verification mainly use news text content to capture the writing differences between fake news and real news [8]. Text-based detection methods are mainly based on the specific language style modeling of fake news, including the early extraction of manual features such as linguistic and thematic features [9, 10]. Castillo et al. [11] proposed a simple model for evaluating the authenticity of Twitter messages by counting the frequency of words, punctuation marks, emoticons, hyperlinks, etc. in the text. Rashkin et al. [12] designed multilingual features using more complex grammatical information with the psycholinguistic feature tool LIWC, and combined it with LSTM networks to construct disinformation recognition models. However, these methods rely on manual design features, which are time-consuming and require specialized domain knowledge, and cannot meet the needs of frequent data processing in the era of big data. The development of deep learning provides a solution for automatic feature extraction, and researchers have been using it to build fake news detection models. Ma

et al. [13] demonstrated the effectiveness of RNN models with word embedding in fake news detection by extracting relevant tweets to form news events. Popat et al. [14] designed an end-to-end speech verification model using news and external evidence statements combined with Bi-LSTM and attention mechanisms.

Whether features of news are manually designed or automatically extracted by deep learning, they can only identify fake information by the text rather than images. Different modal data describing the same news event are often interrelated, which can complement each other. Jin et al. [15] extracted event-related image semantic features by a pre-trained VGG19 model, and used an attention mechanism to extract key information from text and social contexts to adjust the weights of visual semantic features. Experiments show that the method can find many cases of fake news that are difficult to discriminate under a single modality. However, the multi-modal feature representation is still highly dependent on specific events of dataset, which is difficult to migrate and reduces the generalization ability of the model, leading to the failure to identify new events. Therefore, Wang et al. [16] proposed an end-to-end model based on adversarial networks, arguing that the model should be guided to learn event-independent features with more generalization capability. Khattar et al. [17] argue that a simple splicing of text and visual modal features is difficult to adequately express the interaction and association between the two modalities, so they used an encoding-decoding approach to construct a multi-modal feature representation. Singh et al. [18] manually designed text and image features from four dimensions of content, organization, emotion and manipulation respectively, and fused various features through feature cascading to realize fake news detection.

In summary, unimodal such as text-only models have always had limited capabilities for fake news detection. Researches have now started to design discriminative features that can be applied to fake news detection from a cross-modal perspective using a combination of text and images. Multi-modal machine learning, as one of the advantageous methods used by deep learning techniques for feature representation, can exploit the complementarity and circumvent the redundancy between modalities to achieve deep revelation and description of fake news features. However, at present, most of these methods are only applicable to a certain type of fake news images, which are difficult to capture the overall features and represent the complex distribution of image visual contents. Therefore, multi-modal detection of fake news still needs to be further explored to develop a more in-depth multi-modal feature fusion scheme.

3. Data and methodology

3.1. Data description

FACTIFY is the largest multimodal fact verification public dataset consisting of 50K data points, covering news from India and the US [19]. It comes from date-wise tweets from twitter handles of Indian and US news sources: Hindustan Times 1, ANI2 for India and ABC3, CNN 4 for US based on accessibility, popularity and posts per day. From each tweet, the tweet text and the tweet image(s) are extracted. Specifically, FACTIFY contains images, textual claims, reference textual documents and images labeled with five categories. The dataset has a total of 50000 samples, and each of the 5 categories has equal samples with a Train-Val-Test split of 70:15:15. And the descriptions of the labels are as follows:

- Support_Text: the claim text is similar or entailed but images of the document and claim are not similar.
- Support_Multimodal: both the claim text and image are similar to that of the document.
- Insufficient_Text: both text and images of the claim are neither supported nor refuted by the document, although it is possible that the text claim has common words with the document text.
- Insufficient_Multimodal: the claim text is neither supported nor refuted by the document but images are similar to the document.
- Refute: The images and/or text from the claim and document are completely contradictory i.e, the claim is false/fake.

3.2. Unimodal model

First, we use only text and OCR information of images, and propose a model mainly based on combining RoBERTa with other methods (such as adversarial training) and semi-supervised training. Semi-supervised training can further increase the amount of data. At the same time, we obtain the best results by adjusting the weights of different models. The overall architecture of the model is shown in Fig. 1, and the methods used will be described in detail next.

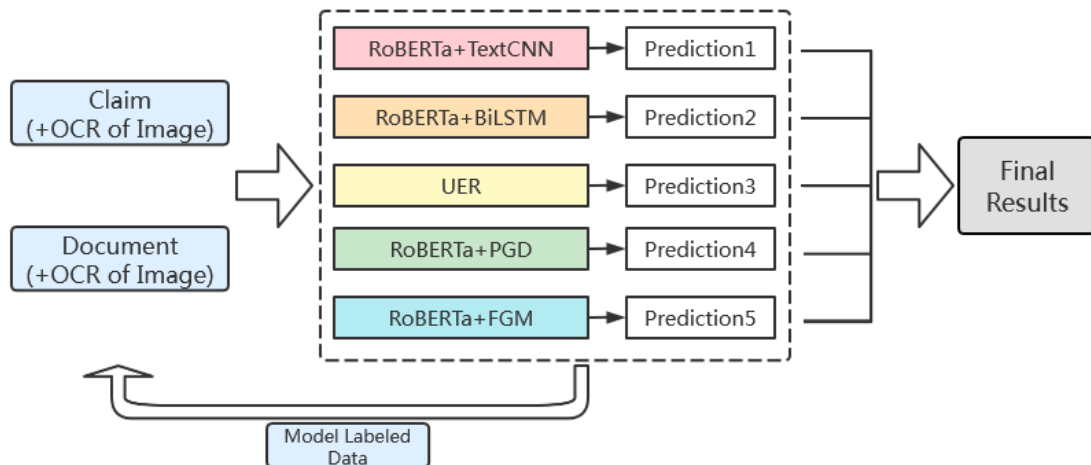


Figure 1: Model structure based only on text-related information.

RoBERTa is an improved version based on Bidirectional Encoder Representation from Transformer (BERT) [20], which achieve significant improvements of BERT on several levels [21]. It mainly made several adjustments based on BERT: 1) Longer training time, larger batch size and more training data; 2) Remove Next Predict Loss; 3) Longer training sequence; 4) Dynamic adjustment of Masking mechanism.

In order to better extract text features for training, we use four methods on the basis of RoBERTa. In TextCNN [22], an embedding representation of the input instance is obtained through an embedding layer, and the characteristics are extracted through a convolution layer.

Finally, a fully connected layer is used to get the final output. We also adopt BiLSTM [23] because it could better capture bidirectional semantic dependencies and facilitate our multi-classification task.

Adversarial training is proposed by Goodfellow in 2015, and this method is also called Fast Gradient Sign Method (FGSM) [24]. FGSM adds a perturbation to the original input instance, and then uses it for training after getting the adversarial sample. Due to its linear characteristics, neural networks are easily attacked by linear disturbances. The adversarial training can improve the robustness of the model when dealing with malicious adversarial samples, and improve the generalization ability. Here we use FGM and PGD of the adversarial training. FGM make a simple modification to the disturbance, canceling the sign function and make a scale in the second normal form [25]. And PGD avoids excessive disturbance by setting a space with a fixed radius [26]. Finally, FGM performs better in our model.

Before UER, there is no pre-trained model that can be perfectly suitable for all tasks, which also brings difficulties to the selection of pre-training models. UER builds an integrated pre-trained toolbox that contains multiple low-coupling modules, which brings the possibility of personalized training tasks [27]. Therefore, in our model, we also try to use the advantages of UER to achieve better results.

Finally, we combine the initial results of the model prediction with the original data to construct a new training dataset, which is known as semi-supervised training [28]. It has been proved that it can obtain better decision boundaries, avoid overfitting, and obtain better performance in our test dataset.

3.3. Multimodal model

In the multi-modal model, we select the last_hidden_state layer of RoBERTa model as the text embedding, and use pre-trained ResNet50 model to extract image features. With the extracted text and image features, we classify them by Light Gradient Boosting Machine (LightGBM) and get the final results. The architecture of multi-modal model is shown in Fig. 2, and the methods involved are described in detail below.

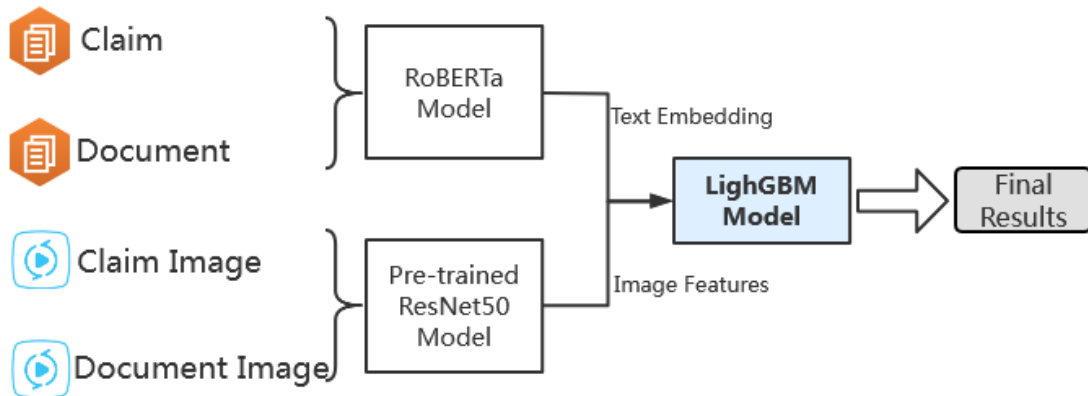


Figure 2: Multi-modal model structure.

With the deepening of the network, the accuracy of the training dataset decreased, which can be determined not to be caused by overfit. To solve this problem, He et al. [29] proposed a new network called Deep residual network (ResNet), which allows the network to be deepened as much as possible (including more hidden layers). It is widely used in target classification and as a part of classical neural Network of computer vision task backbone.

LightGBM is new GBDT implementation with GOSS and EFB[30]. Through their experiments on multiple public datasets, LightGBM speeds up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy. In order to make this method more suitable for our task training, we also combine other methods to learn word vectors including word2vec, fasttext and glove.

4. Experiments and results

First, we only use the text of claim and document, and train 3 epochs with 5e-6 of learning rate and conduct 5-fold cross-validation. We make predictions on different sub-models and get the final results by voting. Next, we combine the text and the OCR of images in claim and document respectively. Here, we also use the unimodal model and find that the model performs better when the image information is combined as shown in Table 1. In our multimodal model, we also use 5-fold cross-validation and the result of multimodal is significantly better than that of unimodal. The best performance of our model ranks the 6th place on the overall test dataset with F1 score 0.7428.

Table 1

The results of our models on test dataset.

Method	Support_Text	Support_Multimodal	Insufficient_Text	Insufficient_Multimodal	Refute	Final
Text	0.6438	0.8290	0.7725	0.7490	1.0	0.6799
Text+OCR	0.7368	0.8506	0.7913	0.7775	0.9977	0.7214
Multimodal	0.7495	0.8602	0.8038	0.8286	0.9913	0.7428

5. Conclusion

In this paper, we propose a new approach to verify fake news, which combines the advantages of various advanced models. The results show that our model performs well in the detecting task, achieving an F1 score of 0.7428. Most importantly, we demonstrate that multimodal models outperform unimodal ones, implying that the adoption of different information does help improve fake news verification. To conclude, the evaluation results indicate that our model is capable of verifying fake news robustly.

The future work can be carried out in three directions: 1) Using unsupervised learning in the data preprocessing stage to solve the problem of data noise; 2) Using transfer learning with attention mechanism to capture important thematic target information in the text and image;

3) Improve the universality of the fake news detection model and extend it to more types of datasets.

References

- [1] P. Meel, D. K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, *Expert Systems with Applications* 153 (2020) 112986.
- [2] P. Heinisch, Stance classification in argument search (2019).
- [3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, *Science* 359 (2018) 1146–1151.
- [4] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, J. Luo, Detection and analysis of 2016 us presidential election related rumors on twitter, in: *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, Springer, 2017, pp. 14–24.
- [5] Edelman, 2016 edelman trust barometer finds global trust inequality is growing (2016).
- [6] Z. Jin, J. Cao, Y. Zhang, J. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification, *IEEE transactions on multimedia* 19 (2016) 598–608.
- [7] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Benchmarking multi-modal entailment for fact verification, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [8] B. Guo, Y. Ding, L. Yao, Y. Liang, Z. Yu, The future of false information detection on social media: New perspectives and trends, *ACM Computing Surveys (CSUR)* 53 (2020) 1–36.
- [9] V. Qazvinian, E. Rosengren, D. Radev, Q. Mei, Rumor has it: Identifying misinformation in microblogs, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1589–1599.
- [10] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, *arXiv preprint arXiv:1708.07104* (2017).
- [11] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.
- [12] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.
- [13] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks (2016).
- [14] K. Popat, S. Mukherjee, A. Yates, G. Weikum, Declare: Debunking fake news and false claims using evidence-aware deep learning, *arXiv preprint arXiv:1809.06416* (2018).
- [15] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.
- [16] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural

- networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.
- [17] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The world wide web conference, 2019, pp. 2915–2921.
- [18] V. K. Singh, I. Ghosh, D. Sonagara, Detecting fake news stories via multimodal analysis, *Journal of the Association for Information Science and Technology* 72 (2021) 3–17.
- [19] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Factify: A multi-modal fact verification dataset, in: Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY), 2022.
- [20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [22] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification, *arXiv preprint arXiv:1510.03820* (2015).
- [23] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), 2016, pp. 207–212.
- [24] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, *arXiv preprint arXiv:1412.6572* (2014).
- [25] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, *arXiv preprint arXiv:1605.07725* (2016).
- [26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, *arXiv preprint arXiv:1706.06083* (2017).
- [27] Z. Zhao, H. Chen, J. Zhang, X. Zhao, T. Liu, W. Lu, X. Chen, H. Deng, Q. Ju, X. Du, Uer: An open-source toolkit for pre-training models, *arXiv preprint arXiv:1909.05658* (2019).
- [28] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, *arXiv preprint arXiv:1610.02242* (2016).
- [29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in neural information processing systems, 2017, pp. 3146–3154.