Effective and Transparent Attributions for Fake News Classification and Search

Marcus Thiel[1], Saijal Shahania[1,2] and Andreas Nürnberger[1]

[1]Otto-von-Guericke-Universität, Universitätsplatz 2, 39106 Magdeburg, Germany
[2]Deutsches Zentrum für Hochschul- und Wissenschaftsforschung, Lange Laube 12, 30159 Hannover, Germany

Abstract

Diverse information sources are becoming more accessible, and news can spread quickly, including fake and highly biased news. It is, therefore, essential to transparently convey to a user what news is potentially fake. Attributing statements inside those fake news to known sources is a potential way to check their validity. However, such attributions are seldom usable due to missing data sets. Therefore, this work aims to define a framework for attributing statements to source documents that do not necessarily contain the exact statements. We are applying our approach to classifying fake news in a search setting and include visual depictions of attributions to explain why something is considered fake or reliable. This method can support users in deciding which news to read and aid in grouping information by credibility.

Keywords

Fake News, Explainability, Attribution, Classification

1. Introduction

Nowadays, technology facilitates more accessible information [1]. Higher accessibility increases the challenge to discern accurate information from falsehoods, particularly with the prevalence of short-form content like TikTok and YouTube shorts [2]. Often, detecting fake news relies on manual efforts and extensive datasets [3]. Attributing statements in a text to a trustworthy source is a good alternative since it increases understanding. However, having an extensive data set is not always doable. Therefore, we propose an approach that balances transparency and classification performance without needing a comprehensive data set. To effectively provide attributions and make them transparent to a user, we need to address three issues: 1. Mapping similar statements in the document to source documents, 2. thresholding statements, so only understandable ones are kept, and 3. using those mappings as attributions for classification.

2. Related Work

Our investigation only assumes the existence of textual data. Hence, we do not consider methods requiring other data like social networks (e.g. Shu, et al. [4]) or metadata. Fake news is commonly detected using stylistic features [5]. Zhou et al. showed that deceiving statements are often more expressive and informal than factual statements. Badaskar et al. identified a few simple syntactical (POS tags, word correlations, etc.) and topic-based (topic correlations) features that were able to achieve relatively high accuracies (91.5%) [6]. Kaliyar et al. are using BERT embeddings with traditional methods like Random Forests and neural networks like an LSTM and a CNN [7]. They almost achieve 99% accuracy on the Kaggle Fake News data set. Some source-based models try to reason over the statements in a particular news article, like in the work by Magdy and Wanas [8] or like the knowledge graph of Shi and Weninger [9]. Both approaches show that source data sets can be small but might not be on par with other methods.

Discovery Science - Late Breaking Contributions 2024
marcus.thiel@ovgu.de (M. Thiel)
0000-0002-9484-1032 (M. Thiel); 0000-0003-1811-6557 (S. Shahania); 0000-0003-4311-0624 (A. Nürnberger)

3. Concept

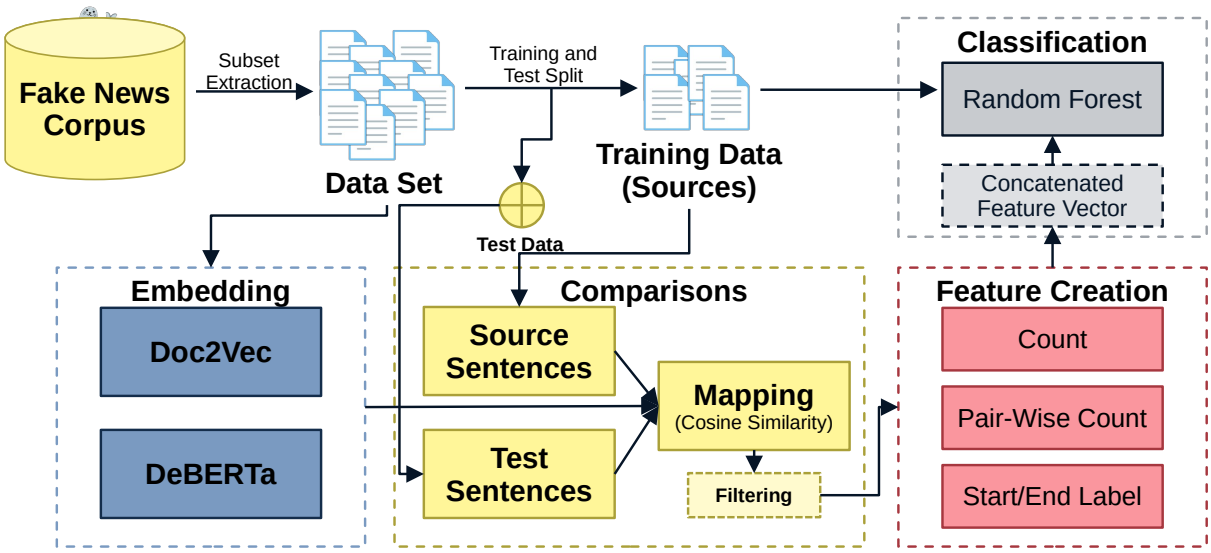

Figure 1: The pipeline for our approach. It shows the three major areas of Embedding, Comparisons and Feature Creation. The Classification part is done using scikit-learn.

Our approach uses deep learning-based semantic embeddings and similarities between sentences. The goal is to support a user's decision on the reliability of an article while maintaining accuracy. The overview of the approach is shown in Fig. 1 and consists of three parts: **1)** *Embedding*: We are using a self-trained Doc2Vec ($|v| = 20$) and a pre-trained DeBERTa called *Base-MNLI*[1]. **2)** *Comparisons*: We calculate the nearest neighbor for sentences using cosine similarity between their embeddings and filter them by thresholding on the average similarity found. **3)** *Feature Creation*: From the nearest neighbors we calculate a feature vector of length 34 containing: (i) the absolute and relative counts of labeled attributions, (ii) the pair-wise counts of two adjacent sentences and (iii) the start and end label of each text. These features are extracted on our two data sets, *KOREA* and *COMPETITION*, which are subsets of the Fake News Corpus[2], filtered by a keyword match where *KOREA* used the keywords *korea* and *nuclear* and *COMPETITION* *fun* and *competition*, leaving 4046 fake and 7917 reliable instances, and 5577 fake and 14996 reliable samples, respectively. We did try known data sets like the Fake News dataset[3], made available by Kaggle and the ISOT Fake News Dataset[4] from the University of Victoria. However, both data sets were not promising since they were easily classifiable using simple stylistic features with over 94% accuracy. We employed ElasticSearch[5] to index the *KOREA* data set and built a simple UI on top, adding a highlighter to show how sentences were attributed in the text. A demo of the UI is shown in a video at https://youtu.be/ZbqgIBQ4cI0.

4. Evaluation

In pre-experiments, we determined that a Random Forest (RF) consistently gave us the best results, which is why we discuss it here. Table 1 shows the results for the test data on the *KOREA* data set. The results for the *COMPETITION* are almost the same. Both data sets indicate a high precision and F1 measure with simple attributions. Only the recall for fake news is not excellent yet. If the threshold of attributions is changed, the recall increases for a high precision cost. However, we concluded that these types of attributions form a robust classifier.

[1] https://github.com/microsoft/DeBERTa
[2] https://github.com/several27/FakeNewsCorpus
[3] https://www.kaggle.com/datasets/jruvika/fake-news-detection
[4] https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php
[5] https://www.elastic.co/elasticsearch

Table 1: Precision, Recall and F1-Score using a RF on the *KOREA* test data.

	precision	**recall**	**f1-score**	**support**
fake	0.95	0.89	0.92	1360
reliable	0.94	0.98	0.96	2588
micro avg	0.94	0.94	0.94	3948
macro avg	0.95	0.93	0.94	3948

Additionally, many attributions provide a good insight into why a text is considered fake. For example, a sourced article describes talks with North Korea about its nuclear program, whereas the target sentence reports a readiness to dismantle it altogether. However, a weakness of the approach is in its nature since some attributions are not sensible from a user's perspective. I.e., the general content of the text is similar, but the sentences have no relationship. The misleading attributions are a challenge that has to be overcome in future work. Many unreliable attributions remain since we optimized the thresholding mostly on classification accuracy. These less sensible attributions can help in classification since they partially stem from a similarly worded article, often from the same source domain. A stricter source selection strategy might help sort out such attributions and decrease run time with only a minor reduction in accuracy.

5. Conclusion

In this paper, we presented a simple method of source-based attribution for fake news classification. This method works reasonably well on a small data set but needs to be tested on a larger domain. Based on initial experiments, the cross-domain accuracy drops significantly due to missing attributions. Hence, more work towards domain-independence and better source selection is required.

References

[1] Y. Kim, Y. Wang, J. Oh, Digital media use and social engagement: How social media and smartphone use influence social activities of college students, Cyberpsychology, Behavior, and Social Networking 19 (2016) 264–269.

[2] A. M. Ostrovsky, J. R. Chen, Tiktok and its role in covid-19 information propagation, Journal of Adolescent Health 67 (2020) 730.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: NAACL-HLT, 2018.

[4] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, dEFEND: Explainable Fake News Detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 395–405.

[5] L. Zhou, J. K. Burgoon, J. F. Nunamaker, D. Twitchell, Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications, Group decision and negotiation 13 (2004) 81–106.

[6] S. Badaskar, S. Agarwal, S. Arora, Identifying Real or Fake Articles: Towards better Language Modeling, in: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II, 2008.

[7] R. K. Kaliyar, A. Goswami, P. Narang, FakeBERT: Fake news detection in social media with a BERT-based deep learning approach, Multimedia Tools and Applications 80 (2021) 11765–11788.

[8] A. Magdy, N. Wanas, Web-based statistical fact checking of textual documents, in: Proceedings of the 2nd international workshop on Search and mining user-generated contents, 2010, pp. 103–110.

[9] B. Shi, T. Weninger, Discriminative predicate path mining for fact checking in knowledge graphs, Knowledge-based systems 104 (2016) 123–133.