UPV-INAOE-Autoritas - Check That: An Approach based on External Sources to Detect Claims Credibility

Bilal Ghanem[1], Manuel Montes-y-Gómez[3],
Francisco Rangel[1,2], and Paolo Rosso[1]

[1] PRHLT Research Center, Universitat Politècnica de València
{bigha@doctor, prosso@dsic, fraranpa@prhlt}.upv.es
[2] Autoritas Consulting, Valencia, Spain
[3] Instituto Nacional de Astrofisica,Óptica y Electrónica (INAOE), Puebla, Mexico
mmontesg@inaoep.mx

Abstract. With the uncontrolled increasing of fake news, untruthful claims, and rumors over the web, recently different approaches have been proposed to address this problem. In this paper, we present a credibility detector of factual claims in presidential debates. Our approach captures the distribution of the results from the search engines to infer the credibility of the claims. We participated in the CLEF-2018 Check That lab for Task 2, obtaining acceptable results.

Keywords: Claims Credibility, English, Arabic.

1 Introduction

A massive amount of information is spread on the web. One of the main disadvantages of this data growth is the uncontrollability of their veracity. The existence of social media networks has helped the increase of untruthful news. Recently, different attempts have addressed this issue to propose solutions for fact checking, for instance for presidential debates. During the debates, multiple claims are made by the candidates about previous facts. Some of these claims could be untruthful: the claimer makes it in an attempt to weaken the other candidate. These untruthful claims pose a real risk on the elections results. In this paper, we present our approach for detecting the factuality or of claims in the presidential debates, where in [1], we presented our approach for task 1 (detecting check worthy claims).

This task concerns with investigating claims veracity in presidential debates. Therefore, a set of presidential debates from the US presidential elections are presented. In our approach, we used results from search engines to infer the veracity of the claims. The idea behind our approach is to infer the veracity by measuring the similarity between the claims and the search engines results, with extracting the results' sources reliability. Also, we modeled the distribution of these features in each search engine. In the following section, we present an overview of the literature. In Section 3, we present our approach, with giving a view on the task. The experiments with the results are presented in Section 4. Finally, in Section 5, we draw some conclusions.

2 Related Work

Researchers in the literature studied several features from different aspects to address this research problem. These features addressed different characteristics of the claims on the web. Authors in [2] have combined two different features to infer the credibility of text. They used language stylistic features to capture the presence of specific language style. Also, they checked the reliability of the text sources using two measures: Amazon AlexaRank and Google PageRank. Authors in [3] used a different way to assess the credibility of claims. In their approach, they created a query from claims using the main sentence components, and they passed it to Google and Bing search engines. From the obtained snippets' results, they trained an LSTM network and they used its encoding to represent the results. These encodings were combined with other similarity features to train two classifiers: a Support Vector Machine (SVM) and a Neural Network. Another set of features were proposed in [4], where the authors studied different aspects of the claims in Twitter to infer their credibility. The authors used features based on the text characteristics, user-based, topic-based, and tweets propagation-based features. In [5] a continuous conditional random field model was used to exploit several signals of interaction between a set of features. In their approach, the authors used features from the language of the news, source trustworthiness, and users confidence. An alternative statement collection approach was proposed in [6]. The authors collected alternative statements of the original claim by changing the doubt unit. Moreover, they inferred the veracity based on different rankers. The only weakness of their approach is that the process is not fully automatic: when a claim is entered, the user should choose the doubt unit which will be investigated. Authors in [7] proposed linguistic, credibility and semantic features to infer the credibility of Bulgarian news. They trained a word2vec model on DBPedia to model the semantics of the documents.

3 Overview of our Approach

As we mentioned, this task concerns with the claims from the presidential debates. The claims that are unworthy for checking have not been annotated and kept in the debates to keep the context. Factual claims have been tagged as True, False, and Half-True. These debates are provided in two languages: English and Arabic, where the Arabic text is translated from the original English debates. The dataset that was provided is imbalanced, where the total number of factual claims is 81; claims as True: 19, Half-True: 22, False: 41. The task goal is to detect the credibility of the provided claims. The macro F1 score was used as the performance measure. More details of the task are mentioned in [10].

Hypothesis: Factual claims have been discussed and mentioned in online news agencies. In our approach, we used the distribution of these claims in the search engines results. Furthermore, we supposed that truthful claims have been mentioned more by trusted web news agencies than the untruthful ones. Therefore, our approach depends on modeling the returned results from search engines using similarity measures and with extracting the reliabilities of the results' sources (dependent features). Also, we captured the distribution of these previous features from each search engine (independent features).

At the beginning, we started to reformulate each claim into a query. We fed this query to the Google and the Bing search engines to obtain a set of results. In our approach, we use only the returned snippets and we do not investigate more the original web pages. Given the search engine results, we used in our approach the first N results for the feature extraction. Next, we built the representation of the features:

Independent features: For each returned result, we extracted the following three features:

> *Cosine over embedding*: we used pre-trained Google News word2vec embedding to measure the cosine similarity between each snippet and the query. We used the main sentence components, discarding the stopwords. In the same way, we built this feature for the Arabic language, but we used fastText pre-trained embedding [8] since Google news word2vec is not available.
>
> *AlexaRank*: For each result, we used Amazon Alexa Rank to retrieve the rank of its site. The sites that have lower values are the sites that have higher reliability.
>
> *Text similarity*: we used another text similarity measure, but using the full sentence components (similarity over tokens), and without text embedding. For the English part, we used the Spacy python library[4], while for the Arabic language, since this library is not available, we implemented the text similarity approach that used in [9] for plagiarism detection.

As we mentioned, we considered the first N results from the search engines, thus, we ended with a features vector of size 3×N.

Dependent features: We extracted a set of features based on the previous independent features. These features model the distribution of the previous feature set, that has been extracted using Average (Avg) and Standard Deviation (Std).

> *Avg and Std of AlexaRank feature*: We computed both Avg and Std features for the Alexa values that were extracted for the first N results.
>
> *Avg and Std of the Cosine over embedding feature*: Similarly, we computed also the Avg and the Std features for the cosine similarities values that were extracted.

[4] https://spacy.io/, visited in May 2018

At the end, our representation has (3×N) + 4 features.

All of these previous dependent and independent features were extracted twice, one from the Google and another one from the Bing search engines. In the following section, we will investigate their importance.

4 Experiments and Results

As we mentioned in the previous section, we built our features based on the first N results from the search engines. Experimentally, we found that choosing the first 5 results (N=5) has produced the highest results. Based on that, our feature vector length is 38 features.

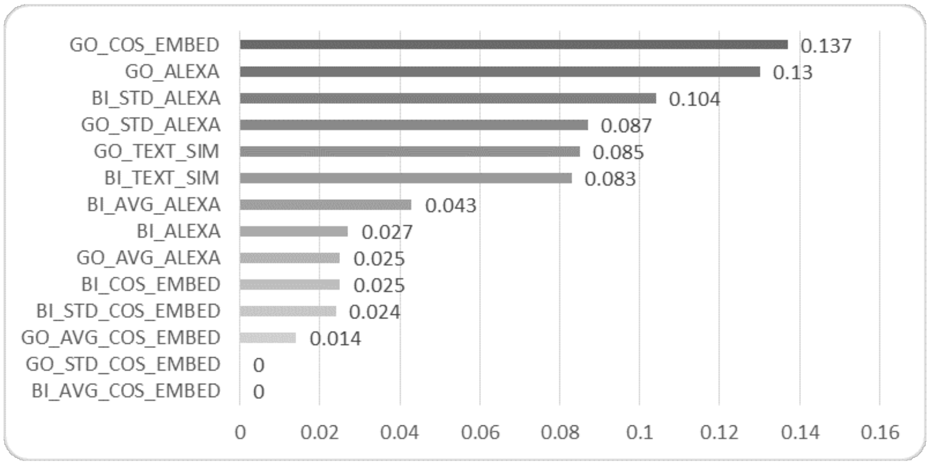

Fig. 1. The Information Gain values of the feature set. The features that started with BI are the ones built using Bing, similarly, GO for Google.

In Figure 1, we show the information gain of these features for each search engine. From these results, we can infer that the features that were obtained by the Google search engine are more important than the Bing features. Based on that, we can notice that the Google results can improve the performance more than the Bing results. At the beginning of our experiments, we tried also to combine Yahoo results, but unfortunately, in all of our experiments, Yahoo results had a lower performance.

Since our approach is search engines-based, for the Arabic task, we found that these claims did not exist because they were written originally in English and translated into Arabic for this task. Therefore, we translated back these claims into English to retrieve results. After that, the results and the query were translated back to Arabic.

During our experiments, many classifiers were tested. We found that the Random Forest classifier achieved the highest results. By using the K-Fold stratified technique, we achieved 0.34 of macro F1 score. The chosen value of K is 5, where we have a small number of data instances and an imbalanced dataset. Thus, higher values of K may lead to absence of some classes in one of the training/testing cycles. We tried also to build a different type of queries, using main sentence components, or phrase queries, but we found that when we changed the query, the results were affected negatively, especially when we used the phrase query, we noticed that the search engines snippets became meaningless (phrases appeared in the snippets but as small text clips connected using "..." characters and combined into the main snippet, where the semantic meaning of the main snippet became biased). For this reason, we passed the queries without any modification, letting the search engines to retrieve the most appropriate results for each one.

For the official testing phase the Mean Absolute Error (MAE) was used as performance measure. In Table 1, the results of the task are shown.

Table 1. Official results for the Task 2, released using the MAE measure.

Team	English	Arabic
Copenhagen	0.705	-
FACTR	0.913	0.657
UPV-INAOE	0.949	0.820
bigIR	0.964	-
Check It Out	0.964	-

In the English part, our approach obtained the 3rd best results, while for the Arabic part, only two teams have submitted their runs. We can observe that the results are low, showing the difficulty of the task.

5 Conclusion

We have presented our simple approach for detecting the veracity of claims in the presidential debates. As we mentioned, our approach uses search engine results to infer the claims veracity. In our approach, we extracted two different types of features, dependent and independent, to model the distribution of claims' results. In general, the results of the task are low, knowing that our approach during the tuning phase has achieved good results comparing to the official one. Also, we can conclude that our feature set has improved the results with respect to the other participants approaches and to the baseline.

6 Acknowledgements

The authors acknowledge the SomEMBED TIN2015-71147-C2-1-P MINECO research project. The work on the data in Arabic as well as this publication were made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the last two authors.

References

1. Ghanem, Bilal, Manuel Montes-y-Gòmez, Francisco Rangel and Paolo Rosso. UPV-INAOE-Autoritas - Check That: Preliminary Approach for Checking Worthiness of Claims. In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CLEF '18, Avignon, France, September.
2. Popat, Kashyap, Subhabrata Mukherjee, Jannik Strtgen, and Gerhard Weikum. Credibility Assessment of Textual Claims on the Web. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173-2178. ACM, 2016.
3. Karadzhov, Georgi, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully Automated Fact Checking Using External Sources. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 344-353. 2017.
4. Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. In Proceedings of the 20th international conference on World wide web, pp. 675-684. ACM, 2011.
5. Mukherjee, Subhabrata, and Gerhard Weikum. Leveraging Joint Interactions for Credibility Analysis in News Communities. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 353-362. ACM, 2015.
6. Li, Xian, Weiyi Meng, and Clement Yu. T-verifier: Verifying Truthfulness of Fact Statements. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on, pp. 63-74. IEEE, 2011.
7. Hardalov, Momchil, Ivan Koychev, and Preslav Nakov. In Search of Credible News. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 172-180. Springer, Cham, 2016.
8. Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. Transactions of the Association of Computational Linguistics 5, no. 1 (2017): 135-146.
9. Ghanem, Bilal, Labib Arafeh, Paolo Rosso, and Fernando Sánchez-Vega. HYPLAG: Hybrid Arabic Text Plagiarism Detection System. In International Conference on Applications of Natural Language to Information Systems, pp. 315-323. Springer, Cham, 2018.
10. Barrón-Cedeño, Alberto and Elsayed, Tamer and Suwaileh, Reem and Màrquez, Lluís and Atanasova, Pepa and Zaghouani, Wajdi and Kyuchukov, Spas and Da San Martino, Giovanni and Nakov, Preslav, Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 2: Factuality. In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CLEF '18, Avignon, France, September.