Using External Knowledge Bases and Coreference Resolution for Detecting Check-Worthy Statements

CLEF-2019 Shared Task: Automatic Identification and Verification of Claims

Salar Mohtaj[1], Tilo Himmelsbach[1],
Vinicius Woloszyn[1], and Sebastian Möller[1,2]

[1] Quality and Usability Lab, Technische Universität Berlin
[2] DFKI Projektbüro Berlin
Berlin, Germany
{salar.mohtaj, tilo.himmelsbach, woloszyn,
sebastian.moeller}@tu-berlin.de

Abstract. With the proliferation of online information sources, it has become more and more difficult to judge the trustworthiness of a statement on the Web. Nevertheless, recent advances in natural language processing allow us to analyze information more objectively according to certain criteria - e.g. whether a proposition is factual or opinative, or even the authority or credibility of an author in a certain topic. In this paper, we formulated a ranking schema that can be employed in textual claims for speeding up the human fact-checking process. Our experiments have shown that our proposed method statistically outperformed the baseline. Additionally, this work describes a multilingual data set of claims collected from several fact-check websites, which was used to fine-tuning our model.

Keywords: fact-checking · check-worthiness · fake news · coreference resolution · political debates

1 Introduction

The 2016 American presidential elections were a source of growing public awareness of what has since been denominated as "fake news". The term started to be used in different positions within the social space as a means of discrediting, attacking and delegitimizing political opponents. However, the task of assessing the credibility of a claim is time-consuming for the user. For example, Kumar's work [12] reports that even humans are not able to always distinguish hoax from

authentic claims and that quite a few people could differentiate satirical articles from true news.

With the increasing number of false claims and rumors, fact-checking websites like *snopes.com*, *politifact.com*, *fullfact.org*, have become popular. These websites compile articles written by experts who manually investigate controversial claims to determine their veracity, providing shreds of evidence for the verdict (e.g. true or false). However, with the quick proliferation of such false statements, especially in the context of a political debate, it becomes very difficult for a single person to assess the validity of all claims made.

In this paper, our team "É Proibido Cochilar"("É Proibido Cochilar" is the title of a Brazilian song and means "it is forbidden to take a nap") have put forward a new supervised worthiness-rank of the checking of a claim. Additionally to the Presidential debates in the 2016 US campaign[3], we have also created a large multilingual data set of statements extracted from several different fact-checking websites. The experiments have shown that our proposed method statistically outperformed the baseline in terms of imitating the human experts judging about the Worthiness of checking a claim.

The remainder of this paper is organized as follows. Section 2 discusses previous works on fake News detection. Section 3 presents details of our approach. Section 4 and 5 describe the design of our experiments and the results. Section 6 summarizes our conclusions and presents future research directions.

2 Related Work

Several studies have addressed the task of assessing the credibility of a claim. For instance, Popat et al. [16] proposed a new approach to identify the **credibility** of a claim in a text. For a certain claim, it retrieves the corresponding articles from News and/or social media and feeds those into a distantly supervised classifier for assessing their credibility. Experiments with claims from the website *snopes.com* and from popular cases of Wikipedia hoaxes demonstrate the viability of Popat et al proposed methods. Another example is TrustRank [9]. This work presents a semi-supervised approach to separate reputable good pages from spam. To discover good pages it relies on an observation that good pages seldom point to bad ones, i.e. people creating good pages have little reason to point to bad pages. Finally, it employs a biased PageRank using this empirical observation to discover other pages that are likely to be good.

Controversial subjects can also be indicative of dispute or debate involving different opinions about the same subject. Detect and alert users when they are reading a controversial web-page is one way to make users aware of the information quality they are consuming. One example of **controversy** detection is [6] which relies on supervised k-nearest-neighbor classification that maps a webpage into a set of neighboring controversial articles extracted from Wikipedia. In this approach, a page adjacent to controversial pages is likely to be controversial itself. Another work in this sense is [13] which aims to generate contrastive summaries of different viewpoints in opinionated texts. It proposes a comparative

LexRank, that relies on random walk formulation to give a score to a sentence based on their difference to other sentences.

Factuality Assessment is another way to asses the information quality. Yu et al.'s work [21] aims to separate opinions from facts, at both the document and sentence level. It uses a Bayesian classifier for discriminating between documents with a preponderance of opinions, such as editorials from regular news stories. The main goal of this approach is to classify a document/sentence in factual or opinionated text from the perspective of the author. The evaluation of the proposed system reported promising results in both document and sentence levels. Other work on the same line is [17], which proposes a two-stage framework to extract opinionated sentences from news articles. In the first stage, a supervised learning model gives a score to each sentence based on the probability of the sentence to be opinionated. In the second stage, it uses these probabilities within the HITS schema to treat the opinionated sentences as Hubs, and the facts around these opinions are treated as the Authorities. The proposed method extracts opinions, grouping them with supporting facts as well as other supporting opinions.

There are also some works that analyze how a piece of information flows over the internet. For instance, [7] presents an interesting analysis about how Twitter bots can send spam tweets, manipulate public opinion and use them for online fraud. It reports the discovery of the 'Star Wars' botnet on Twitter, which consists of more than 350,000 bots tweeting random quotations exclusively from Star Wars novels. It analyzes and reveals rich details on how the botnet is designed and gives insights on how to detect **virality** in Twitter.

Other works analyze the writing style in order to detect a false claim. [10] reports that fake news in most cases are more similar to satire than to real news, leading us to conclude that persuasion in the fake news is achieved through heuristics rather than the strength of arguments. It shows that the overall title structure and the use of proper nouns in titles are very significant in differentiating fake from real. It gives an idea that fake news is targeted for audiences who are not likely to read beyond titles and that they aim at creating mental associations between entities and claims. Decrease the **readability** of texts is also another way to overshadow false claims on the internet. Many automatic methods to evaluate the readability of texts have been proposed. For instance, Coh-Metrix [8], which is a computational tool that measures cohesion, discourse, and text difficulty.

Most of the works just cited rely on supervised learning strategies addressed to assess News articles using few different aspects, such as credibility, controversy, factuality and virality of information. Nonetheless, a common drawback of supervised learning approaches is that the quality of the results is heavily influenced by the availability of a large, domain-dependent annotated corpus to train the model. Unsupervised and semi-supervised learning techniques, on the other hand, are attractive because they do not imply the cost of corpus annotation. In short, our method uses a semi-supervised strategy where only a small

set of unreliable News websites is used to spot another bad News websites using a biased PageRank.

3 Proposed Approach

In order to rank statements according to their estimated check-worthiness, we relied on an important empirical observation: there is a significant number of claims with pronouns referring back to nouns mentioned in previous statements. For example, "I beat her, and I beat her badly. She's raising your taxes really high"; the pronouns *her* and *she* refer to the same person, namely Hillary Clinton. More examples are given in table 1.

Table 1. Sample sentences from the training data that contain pronouns referring back to nouns mentioned in previous statements.

Speaker	**Sentence**	**Label**
SANDERS	They are working longer hours for low wages.	1
TRUMP	I beat her, and I beat her badly.	1
CLINTON	They're interested in keeping Assad in power.	0
SANDERS	Listen to what I told them then.	0
TRUMP	She's raising your taxes really high.	1

Sentences that contain pronouns are normally an issue for statistical models and can significantly decrease the quality of prediction. To overcome this issue, a coreference resolution technique is applied to replace pronouns with their original references. We used a feed-forward neural-network to compute the coreference score for each pair of potential mentions [1], e.g. *Hillary Clinton* ← *she*. We have considered the last 30 sentences (slide-window) to compute the coreferences. Table 2 illustrates the coreference resolution of the examples presented in Table 1. To resolve coreferences leads to more clear-cut statements, which in our experiments improved the performance of our predictions.

Table 2. The result of applying coreference resolution on the sentences in Table 1.

Speaker	**Sentence**	**Label**
SANDERS	Millions of Americans are working longer hours for low wages.	1
TRUMP	I beat Hillary Clinton, and I beat Hillary Clinton badly.	1
CLINTON	Russia is interested in keeping Bashar al-Assad in power.	0
SANDERS	Listen to what I told YouTube then.	0
TRUMP	Hillary Clinton's raising your taxes really high.	1

Additionally, we have performed a normalization of the corpus using standard techniques: lowercasing, lemmatization, number removal, white-space removal,

stop-word removal, and tokenization. In addition to preparing the data set to the training phase, we used some external fact-checking collection to tackle some issues in the provided data set. Firstly, since the provided data is highly imbalanced (less than 3% of data are labeled as 1), we provide external data to make the data more balanced. Moreover, it can lead to an improved generalization of the classification model if the training data is more diverse. To add the external data-set to the training data same pre-processing steps includes in coreference resolution, are applied on the data.

For this purpose, we have created a tool - called **Fake News Extractor**[2] - to automatically extract claims from Fact-Checking websites and then consolidate a large data set for machine learning purposes. It extracts claims in three different languages: English, Portuguese and German. Table 3 gives some statistics about the data set created by our tool.

Table 3. Claims used to train our model.

URL	Language	#
http://fullfact.org	English	27594
http://www.snopes.com	English	
http://www.politifact.com	English	
http://TruthOrFiction.com	English	
http://checkyourfact.com	English	
http://piaui.folha.uol.com.br/lupa/	Portuguese	1463
http://aosfatos.org/aos-fatos-e-noticia/	Portuguese	
http://apublica.org/checagem/	Portuguese	
http://g1.globo.com/e-ou-nao-e/	Portuguese	
http://www.e-farsas.com/	Portuguese	
http://www.mimikama.at/	German	5193
http://correctiv.org/	German	

We have used Support Vector Machine Regression (SVM) [18] and Term Frequency–Inverse Document Frequency (TF-IDF). Additionally, we have used Scikit-Learn [14] library for feature extracting, for example uni-gram, bi-grams and tri-grams. In a nutshell, the main contributions to tackle the challenge are as follows:

- the use of using coreference resolution in political debates
- creation of an external collection of claims extracted from fact-check websites employed as a training set

4 Experiment Design

For the validation of our experiments, we used 5-fold cross validation in the document level. In other words, we have splited the training data into 5 categories, where each fold of the whole document is considered as belonging to either the

training or testing set. The reason for splitting the data into training and testing folds in the document level is to preserve the sequence of sentences of each debate.

We have created three different models, as follow:

Resolving coreference (ReCo): we have tested the performance of our model using the normalization of the corpus - previously described in Section 3.

Resolving coreference + further pre-processing (ReCo+pre): as described in the previous Section, in this experiment the coreference resolution technique is used to replace pronouns by the right references. We also employed in this model the normalization of the corpus.

Using external fact-checking data-set (ExtDat): in this model we used an external data set of claims described previously. Additionally, all mentioned text normalization techniques were used in this experiment.

5 Results

In this section, we present the results and discuss the evaluation of our proposed approach for Worthiness-Rank of Claims Checking.

Figure 1 shows that our models yield better results in comparison to the baseline. The differences range from 4.02 to 8.86 percentage points (pp) when compared to the runner-up method, namely *ExtDat*. Using a Wilcoxon statistical test [19] with a significance level of 0.05, we verified that the results of our models are statistically superior to the baseline.

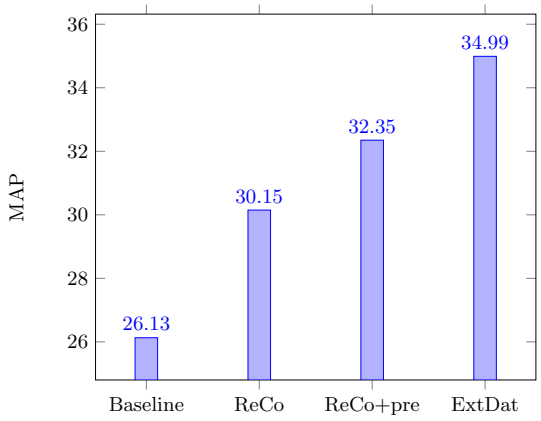

Fig. 1. The obtained results in each experiment in addition to baseline results

Regarding the final submission, we used the 2-top best models, namely *ReCo+pre* and *ExtDat* models as our contrastive and primary submissions, respectively. Table 4 presents our results on the test data in different evaluation measures.

Table 4. Our primary and contrastive results on the test data

Submission	MAP	RR	R-P	P@1	P@3	P@5	P@10	P@20	P@50
Primary	.079	.351	.088	.142	.238	.142	.128	.107	.0714
Contrastive	.135	.541	.159	.428	.238	.257	.271	.164	.120

6 Conclusions and Future Work

The performance of a machine learning model trained in a supervised manner is mostly determined by the amount and quality of the training data. The paradigm of transfer-learning can be a remedy to the problem of having only small amounts of human-labeled data [11]. Language models that are trained unsupervised on a large but unlabeled corpus from a similar domain tend to learn abstract/high-level features that can benefit supervised training [15]. We assume that the basic understanding of a language that is learning by Language Models like ELMo [15], XLNet [20], and BERT [5] can be of particular use for teaching the machine the concept of check-worthiness. Furthermore, check-worthiness could be interpreted as more than a pure language understanding problem. The overall goal of reducing the human workload of checking claims could be further approached by a Fact-Checking system based on the ideas of question answering over knowledge-bases [4]. This way obvious true or false claims could be filtered out. Factual claims like "Homicides last year increased by 17 percent in America's fifty largest cities." are relatively easy to verify compared to "[...] NAFTA [is] one of the worst economic deals ever made by our country.".

References

1. Github - huggingface/neuralcoref: Fast coreference resolution in spacy with neural networks. https://github.com/huggingface/neuralcoref, accessed: 2019-06-30
2. Github - vwoloszyn/fake_news_extractor: This project is a collective effort to automatically extract claims from fact-checking websites and then consolidate a large data set for machine learning purposes. currently, these claims are available for english, portuguese and german. https://github.com/vwoloszyn/fake_news_extractor, accessed: 2019-06-25
3. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G.: Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness
4. Berant, J., Chou, A., Frostig, R., Liang, P.: Semantic parsing on Freebase from question-answer pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1533–1544. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dori-Hacohen, S., Allan, J.: Detecting controversy on the web. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1845–1848. ACM (2013)

7. Echeverria, J., Zhou, S.: Discovery, retrieval, and analysis of the'star wars' botnet in twitter. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 1–8. ACM (2017)
8. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix: Providing multilevel analyses of text characteristics. Educational researcher **40**(5), 223–234 (2011)
9. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. pp. 576–587. VLDB Endowment (2004)
10. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398 (2017)
11. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
12. Kumar, S., West, R., Leskovec, J.: Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of the 25th International Conference on World Wide Web. pp. 591–602. International World Wide Web Conferences Steering Committee (2016)
13. Paul, M.J., Zhai, C., Girju, R.: Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 66–76. Association for Computational Linguistics (2010)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
15. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
16. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: 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)
17. Rajkumar, P., Desai, S., Ganguly, N., Goyal, P.: A novel two-stage framework for extracting opinionated sentences from news articles. In: Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing. pp. 25–33 (2014)
18. Vapnik, V.: The Support Vector Method of Function Estimation, pp. 55–85. Springer US, Boston, MA (1998). https://doi.org/10.1007/978-1-4615-5703-63, https://doi.org/10.1007/978-1-4615-5703-63
19. Wilcoxon, F., Katti, S., Wilcox, R.A.: Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. Selected tables in mathematical statistics **1**, 171–259 (1970)
20. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)
21. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. pp. 129–136. Association for Computational Linguistics (2003)