A Framework for Automated Fact-Checking for Real-Time Validation of Emerging Claims on the Web

Andreas Hanselowski and Iryna Gurevych

Research Training Group AIPHES,
Technische Universität Darmstadt

October 23, 2017

Literally minutes after the Las Vegas shooting, rumors about the identity and the motives of the perpetrator started to circulate on Twitter. Some of the most persistent ones were that the perpetrator was a Muslim convert [1] or a member of Antifa [2]. Moreover, also alleged reactions of people to the shooting received much attention. On a fake Twitter account, a liberal teacher was supposedly hoping that Trump supports were among the victims ([3]). When all the attention is on the event and not all details about the incident are known, propagandists have the opportunity to instrumentalize the event to promote a certain worldview.

As the problem of false information being distributed on the web became more severe in the past couple of years, there is also an increased interest in information validation. Thus, fact-checking websites like politifact.com, fullfact.org, and snopes.com are becoming more popular. On these websites, journalists or professional fact-checkers are manually resolving controversial claims, by providing a verdict, which is backed up by evidence. Nevertheless, even though manual fact-checking blossoms from the spread of fake news, the approach is rather mitigating the influence of false information rather than solving the problem. The resolution is often done subsequently after a fake news article has spread, although most of the damage is caused when the fake news article is distributed through social networks. In fact, many of the news consumers are not going to review the facts on a story once the spotlight of the media has shifted to a different topic. Thus, real-time fact-checking techniques are required, which would be able to intervene in the proliferation process in the early stages, before the false information goes viral.

Many of the issues of manual claim validation can be addressed by automated fact-checking, as it would be possible to validate a large number of articles as they appear on the web automatically. To address the problem, a number of different approaches have been suggested, many of which are based on knowledge bases ([4, 5, 6]). These methods are validating a claim by verifying whether it is consistent with the knowledge base, that is, whether predicates can be found which basically restate the claim or contradict it. Nevertheless, knowledge bases only represent a small portion of all the information available on the web and newly generated knowledge is rare since the updating process requires some time. Thus, in particular for real-time fact-checking, methods based on raw text are more suitable, as they would allow recently published web documents to be incorporated into the validation process. However, claim validation on the basis of raw text has not yet received much attention and only a few studies address these issue.

The task 8 in SemEval-2017 [7] was concerned with the problem of validating claims on Twitter. The claim itself was represented by a tweet and the problem was approach in two different settings: In the closed setting, the validation was done only on the basis of the features of the claim tweet itself. In the open setting, external information, in form of related Wikipedia articles and web documents, was provided. Both problem settings turned out to be too challenging for the applied methods since the majority baseline could not be beaten. The participants suggest implementing additional more discriminative features, like those used in sentiment analysis, or discriminative rules to further increase performance.

A method for the identification and validation of simple claims about 16 statistical properties of countries is presented in [8]. The authors introduce a distantly supervised approach which is based on a knowledge base, as well as raw text input. The method is able to identify statistical claims with 60% precision and to validate these claims without explicit supervision.

The framework for claim validation presented in [9] is to our knowledge the most comprehensive. The authors extracted 4856 claims and the verdicts for these claim from the fact-checking website snopes.com [10]. In order to collect external information for the resolution of the claims, the Google search engine was used. The developed system is able to determine the stance of a text with respect to a given claim, the credibility of sources and the validity of the claim. The authors report 80% accuracy for the claim validation task.

Nevertheless, despite significant progress in the field of natural language processing in the past couple of years, a fully automated system for claim validation, which is able to validate newly emerging claims on the web with high accuracy, is not yet feasible. Today's approaches for automated fact-checking are still restricted in their capabilities and are only trained on small amounts of data. The validation process is very challenging and there are a number of abilities a system must have, such as the incorporation of world knowledge in the validation process or the ability to reason with known facts, which cannot be easily realized with today's machine learning techniques. Thus, our objective, is therefore, to develop a system, which is able to assist a fact-checker in the validation process in order to speed up the procedure rather than taking over the job entirely.

In order to address the described challenges, we are proposing a comprehensive system for claim validation which has the following characteristics. For the reduction of the complexity of the problem, we divide the task into several subproblems and tackle them individually. As a result, also the transparency of the system is increased, which enables the fact-checker to comprehend why a particular verdict was predicted on the basis of the intermediate outputs of the subsystems. To address the problem of data sparsity in knowledge bases, we are developing a system which extracts its knowledge from web documents. This would enable the system to assess the veracity of a claim on a wide range of topics.

The pipeline of the proposed system is illustrated in Figure 1. In the first step, relevant web documents for the resolution of a given claim, as well as the information about their sources, are retrieved. In the second step, evidence, which supports or refute the claims, is identified in the web documents. The stance of the evidence with respect to the claim is determined in the third step. In the fourth step, the actual claim validation is performed. The generated output of all three previous subsystems serves thereby as an input.

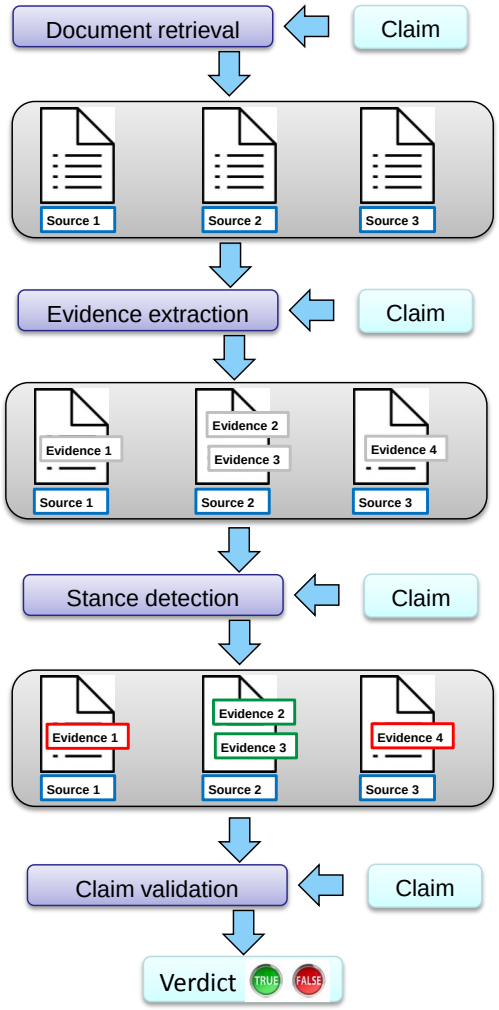

Figure 1: Pipeline for claim validation

Since our objective is to develop a system for automated fact-checking, which is transparent, the identification of evidence in the validation process is one of the main contributions. We are planning to find text snippets which are crucial for the interpretation of the verdict by the fact-checker, as well as for the machine learning model which comes up with the verdict.

The development of the system for claim validation is currently in progress, and we have already implemented methods for evidence extraction, stance classification, and claim validation. The machine learning methods are trained on a corpus, which was constructed by crawling the snopes.com website [10]. In contrast to the study presented in [9], in addition to the claims and the verdicts, we have also collected evidence for each claim from the Snopes website, and the documents, from which the evidence have been extracted.

Evidence extraction for automated claim validation is considered as a classification problem on the sentence level. We have found that for this task, feature-based classifiers, such as linear models and SVMs, outperform neural networks based on LSTMs ([11]). Nevertheless, even the linear model, which performed best, reached a relatively low F1 score of 55%. We believe that the low performance is due to a low upper bound for the task since the fact-checkers have not ensured that their annotations are reproducible.

For stance detection, a feature-based multilayer perceptron ([12]) was used, which was one of the best performing models in the Fake News Challenge stance detection task [13]. We have implemented additional features for the model and have been able to increase the performance from 81.97% to 82.7% on the Fake News Challenge evaluation metric.

For the claim validation, different LSTM network structures have been applied. We have found that regular BiLSTM and hierarchical BiLSTM models [14] perform well for the task and even outperform BiLSTM models with different kinds of attention. The highest F1 score of 66% was reached by the BiLSTM model.

References

[1] Linda Qiu. False isis connections, nonexistent victims and other misinformation in the wake of las vegas shooting. *The New York Times*, October 2017.

[2] Abby Ohlheiser. A running list of viral hoaxes and misinformation about the las vegas shooting. *The Washington Post*, October 2017.

[3] Abby Ohlheiser. A 'liberal teacher' became a conservative enemy for a viral vegas tweet, but does she actually exist? *The Washington Post*, October 2017.

[4] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[5] Baoxu Shi and Tim Weninger. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133, 2016.

[6] Baoxu Shi and Tim Weninger. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 101–102. International World Wide Web Conferences Steering Committee, 2016.

[7] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[8] Andreas Vlachos and Sebastian Riedel. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601. Association for Computational Linguistics, 2015.

[9] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee, 2017.

[10] Snopes: The urban legends reference pages. `http://www.snopes.com/`. Accessed: 2017-10-16.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Andreas Hanselowski, Avinesh P.V.S, Benjamin Schiller, and Felix Caspelherr. Description of the system developed by team athene in the fnc-1. `https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf`.

[13] The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. `http://www.fakenewschallenge.org/`. Accessed: 2017-10-20.

[14] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, 2016.