Propositional Claim Detection

Keywords: Claim Detection, Disinformation, Fact-Checking, NLP, Data Quality

Extended Abstract

INTRODUCTION Extracting meaning from text is a central part of Communication Science and in the recent years this is increasingly done with the aid of computational methods (van Atteveldt and Peng 2018). Especially fact-checking and research on misinformation profit strongly from this trend. And as large-scale studies on misinformation become more frequent, so does the need for fitting tools. This contribution aims at extending this toolkit. An annotated German-language data set and an NLP-model for Propositional Claim Detection (PCD) are introduced.[1] PCD is the task of classifying claims that have propositional content, i.e., claims that can be true or false. PCD is compatible with various other computational methods and research agendas, for the present purposes, however, it is meant as a foundational building block for the automated detection of claims that carry misinformation. For assuring best results, a variety of methods for increasing data quality were applied. The resulting classifier achieves an F_1-score of 94% and the experiments indicate that it adapts well across domains.

PREVIOUS WORK Claim Detection is usually modelled as a binary text classification task to identify (non-) checkworthy claims. Annotators are asked to code checkworthiness based on if the claim is interesting to the general public (Arslan et al. 2020) or harmful to society (Firoj et al. 2022). However, this definition has been criticized before for being too dependent on prior knowledge of the annotators (Allein and Moens 2020). Moreover, there is not necessarily a unique definition to checkworthiness, as studies found that different fact-checking organizations do not agree on which claims require checking (e.g., Lim 2018). PCD leverages mostly grammatical information to identify claims. This makes the tool less prone to individual background knowledge and ambiguity.

ANNOTATION PROCESS For creating a diverse data set that spans across multiple domains, single sentences from (German-language) newspapers, TV talk shows, plenary protocols, party manifestos, and Tweets were sampled. The annotation was performed by 4 trained annotators (Krippendorff $\alpha = 0.72$). The data set has four labels: *Assertion*, *Opinion*, *Prediction*, and *Other*. Sentences that can be true or false are almost exclusively declarative sentences. Other sentence types like questions or imperatives were labelled as *Other*. However, not all declarative sentences are relevant to fact-checking and misinformation. *Opinions* express subjective world views that do not have to be grounded in facts. *Predictions* are declarative sentences in the future tense that cannot be verified in the present. In PCC the class, which is most relevant, is *Assertion*. This class contains declarative sentences in the present or past tense that do not express personal preferences or world views. Our assumption is that claims that carry misinformation are in almost all cases *Assertions*.

CORPUS STATISTICS Fig. 1 displays the most relevant features of the data set. For assuring high quality, CROWDLAB (Goh et al. 2022) was used to assist the annotators. With this method, an ensemble of machine learning classifiers is used to identify the correct label in case of disagreement between the annotators. Furthermore, Confident Learning was used to remove noisy labels (Northcutt et al. 2021). Each configuration of the data set was tested as a separate experiment (E0-4). For each experiment, we used a different strategy to derive the final label of a sentence. For E0 a strict majority vote was performed, in which only sentences with an agreement of at least 3 annotators are included. E1 is a soft majority vote, for which all

[1] The code can be found on the author's Github page and the data set will soon be published.

sentences were included and in case of disagreement one of the competing labels was chosen. For E2, the label errors of E1 were discarded, using Confident Learning. For E3, labels were chosen using CROWDLAB and for E4 the label errors from E4 were discarded. For all experiments *Assertion* occurs most often, followed by *Other*, *Opinion*, and *Prediction*.

EVALUATION In order to test the data, rather than the models, a broad variety of architectures and embedding techniques was applied. Fig. 2 displays the scores for all experiments. While more traditional models, like SVM, remain in the area of 70+% in F_1, transformer models, independent of the exact architecture, achieve 90+% in F_1 with a maximum of 94%. Across all configurations, there is a pattern that shows that E4 yields the best results. The difference in scores between different experiments can be as high as 16%. Removing noisy labels has the strongest impact. This highlights the importance of data quality. For testing domain adaptation, the data set was split into train and test set according to different criteria, e.g., the topic, media type, or time period (Fig. 3). Experiments show that models that have not seen a certain text topic, media type, or time period during training, still reach scores of 90+% in F_1 when tested on the new domain. Even for the notoriously difficult adaptation to social media data, models still achieve an F_1 score of 85%

CONCLUSION In this contribution, a data set and multiple models for PCD were introduced. PCD's design is less vague than other approaches to claim detection because it focuses stronger on grammatical information. The results show solid performance on different configurations of the data set and across domains. By design, PCD is not sufficient to spot checkworthy claims as it lacks a criterion of relevance in order to prioritize some claims over others. However, this makes it compatible with other computational tools. A possible line of future research is to enhance PCD with further heuristics such as topic models or the automated detection of news values. The resulting tool could be used to identify substantial claims in various contexts.

References

Allein, L., & Moens, M.-F. (2020). Checkworthiness in Automatic Claim Detection Models: Definitions and Analysis of Datasets. In M. van Duijn, M. Preuss, V. Spaiser, F. Takes, & S. Verberne (Eds.), *Disinformation in Open Online Media* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-61841-4_1

Arslan, F., Hassan, N., Li, C., & Tremayne, M. (2020). A Benchmark Dataset of Check-Worthy Factual Claims. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*.

Firoj, A., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Martino, G. D. S., Abdelali, A., Sajjad, H., Darwish, K., & Nakov, P. (2021). COVID-19 Infodemic Twitter dataset. *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)*. https://doi.org/10.7910/DVN/XYK2UE

Goh, H. W., Tkachenko, U., & Mueller, J. (2022). *Utilizing supervised models to infer consensus labels and their quality from data with multiple annotators*. arXiv. https://doi.org/10.48550/arXiv.2210.06812

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, *5*(3). https://doi.org/10.1177/2053168018786848

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research*, *70*. https://doi.org/10.1613/jair.1.12125

Settles, B. (2010). *Active Learning Literature Survey* (No. 1648). University of Wisconsin.

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3). https://doi.org/10.1080/19312458.2018.1458084

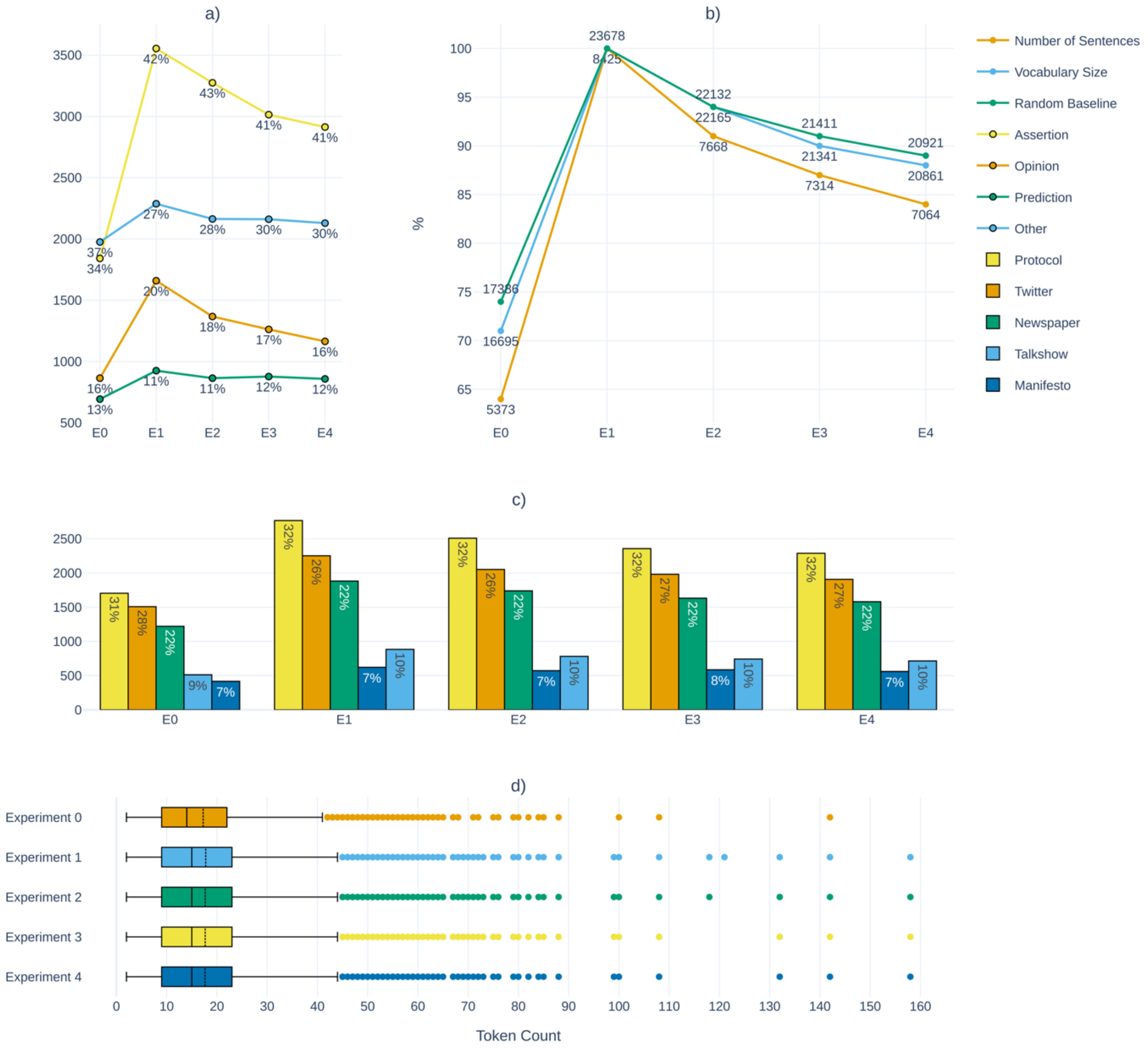

Fig 1 Corpus Statistics for the different experiments. a) Label distribution, b) number of sentences and vocabulary size, c) media type distribution, and d) token count per sentence.

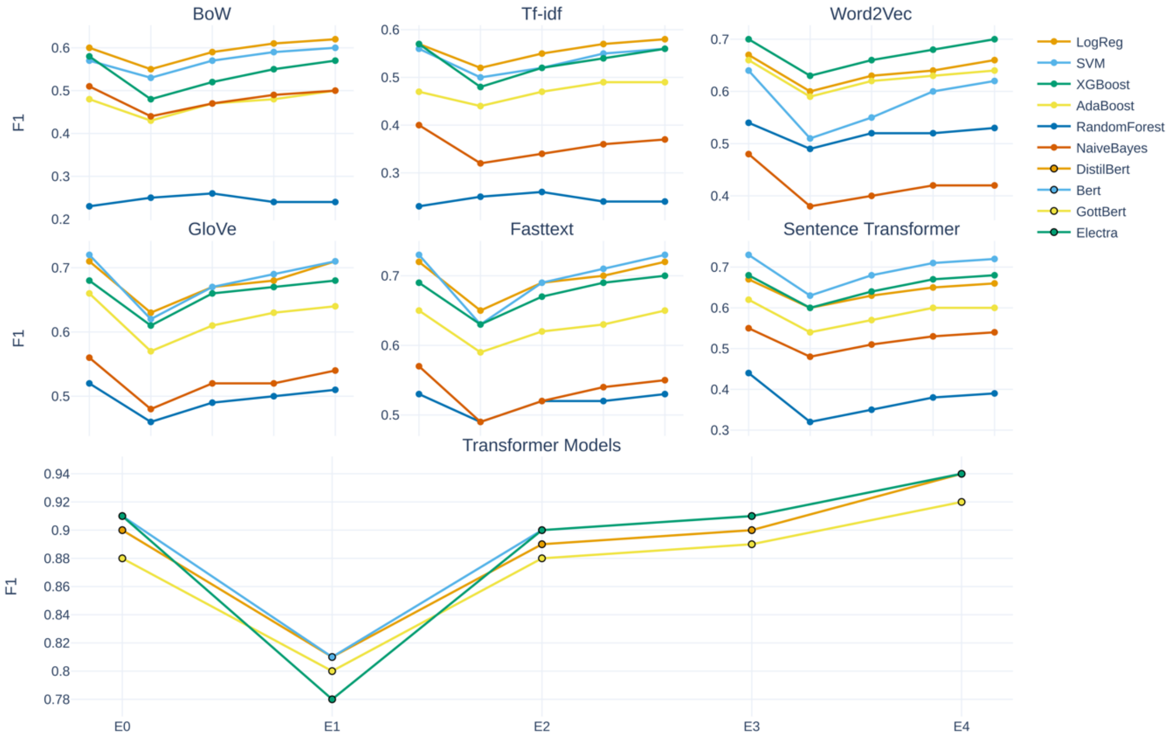

Fig 2 F$_1$ Scores for the different experiments.

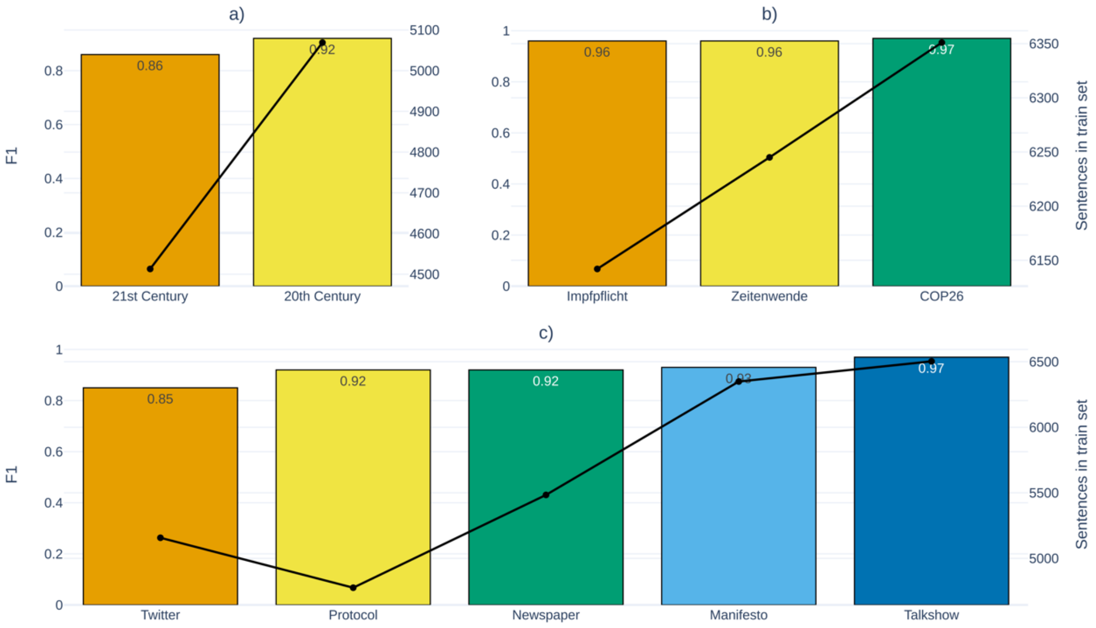

Fig 3 Domain adaptation. a) F_1 Scores when trained/tested exclusively on sentences from the 20th/21st century, b) F_1 Scores when tested exclusively on sentences related to “Impfpflicht” (mandatory vaccination), “Zeitenwende” (Chancellor Olaf Scholz’ speech at the German Bundestag when Russa attacked Ukraine in February 2022), or “COP 26” (Climate Summit in Glasgow) and trained on the rest, c) F_1 Scores when tested exclusively on sentences from Twitter, Plenary Protocols, Newspaper articles, party Manifestos, or political TV talk shows and trained on the rest.