1 Allgemeine Angaben

DFG-Geschäftszeichen: *KL 2869/5-1*
Projektnummer: *667374*
Titel des Projekts: *Automatic Fact Checking for Biomedical Information in Social Media and Scientific Literature (FIBISS)*
Name(n) des/r Antragstellenden: *Prof. Dr. Roman Klinger*
Dienstanschrift/en: *Otto-Friedrich-Universität Bamberg, Lehrstuhl für Grundlagen der Sprachverarbeitung, 96045 Bamberg, Germany*
Name(n) der Mitverantwortlichen:
Name(n) der Kooperationspartnerinnen und –partner:
Berichtszeitraum (gesamte Förderdauer): 01.02.2021–31.07.2024

2 Summary

Zusammenfassung

Die Erforschung von Methoden zur automatischen Überprüfung von Fakten, also Computermodelle, welche korrekte Information von Fehlinformation oder Desinformation unterscheiden können, fokussiert weitestgehend auf die Nachrichtendomäne sowie auf die Analyse von Beiträgen in sozialen Medien. Hierbei werden unter anderem Texte auf ihren Wahrheitsgehalt geprüft. Dies kann durch die Analyse von linguistischen Merkmalen geschehen, die auf eine Täuschungsabsicht schließen lassen, oder durch einen Abgleich mit anderen Quellen, die inhaltlich vergleichbare Aussagen tätigen. Die meisten Arbeiten legen den Schwerpunkt hierbei auf politisch relevante Bereiche.
Ein Gebiet mit besonderer gesellschaftlicher Relevanz ist aber auch die biomedizinische Domäne. In sozialen Medien teilen verschiedene Akteure und medizinische Laien Berichte zu Behandlungmethoden, Erfolgen und Misserfolgen, wie zum Beispiel die (widerlegte) Methode, Virusinfektionen mit Entwurmungsmitteln oder Desinfektionsmitteln zu behandeln. Es finden sich auch Berichte zu (widerlegten) Zusammenhängen zwischen Behandlungen und unerwünschten Wirkungen, wie zum Beispiel die Verursachung von Autismus durch Impfungen.
Die biomedizinische Domäne profitiert allerdings, im Gegensatz zu anderen für die automatische Faktenüberprüfung relevanten Bereichen, von einer großen Ressource verlässlicher wissenschaftlicher Artikel. Das Ziel des Projekts FIBISS war es daher, Methoden zu entwickeln und zu evaluieren, welche biomedizinische Behauptungen in sozialen Medien extrahieren können und diese mit verlässlichen Quellen abgleicht. Eine Herausforderung ist hierbei, dass in sozialen Medien typischerweise keine Fachsprache verwendet wird, so dass unterschiedliche Vokabularien miteinander verbunden werden müssen. Der Ansatz in FIBISS war daher, generalisierende Informationsextraktionsmethoden zu entwickeln. Im Verlauf des Projekts haben

sich zusätzlich große Sprachmodelle prominent als weiterer methodischer Ansatz platziert. Das Projekt wurde daher im Verlauf dahingehend angepasst, generelle Repräsentationen von Behauptungen so zu optimieren, dass sie für den Vergleich mit Hilfe automatischer Fakten-Überprüfungsverfahren geeignet sind.
Im Ergebnis tragen wir Textkorpora bei, die zur Entwicklung und Evaluierung von Systemen zur automatischen biomedizinischen Faktenüberprüfung eingesetzt werden. Wir schlagen Methoden vor, die automatisch Behauptungen so umformulieren, dass sie geeignet sind, automatisch überprüft zu werden. Des Weiteren präsentieren wir Ansätze, die automatisch die Glaubwürdigkeit von Aussagen, auch unabhängig von vorhandener Evidenz, abschätzen können.

Summary

Research into methods for the automatic verification of facts, i.e., computational models that can distinguish correct information from misinformation or disinformation, is largely focused on the news domain and on the analysis of posts in social media. Among other things, texts are checked for their truthfulness. This can be done by analyzing linguistic features that suggest an intention to deceive or by comparing them with other sources that make comparable statements in terms of content. Most studies focus on politically relevant areas.
The biomedical domain is also an area of particular social relevance. In social media, various actors and medical laypersons share reports on treatment methods, successes and failures, such as the (disproven) method of treating viral infections with deworming agents or disinfectants. There are also reports on (disproven) links between treatments and adverse effects, such as the causation of autism by vaccination.
However, the biomedical domain, unlike other areas relevant for automated fact checking, benefits from a large resource of reliable scientific articles. The aim of the FIBISS project was therefore to develop and evaluate methods that can extract biomedical claims in social media and compare them with reliable sources. One challenge here is that social media does not typically use technical language, so different vocabularies have to be combined. The approach in FIBISS was therefore to develop generalizing information extraction methods. In the course of the project, large language models also became prominent as a further methodological approach. The project was therefore adapted to optimize general representations of claims in such a way that they are suitable for comparison using automatic fact-checking procedures.
As a result, we contribute text corpora that are used to develop and evaluate automated biomedical fact-checking systems. We propose methods that automatically reformulate claims so that they are suitable to be automatically verified. Furthermore, we present approaches that can automatically assess the credibility of claims, even independently of existing evidence.

3 Wissenschaftlicher Arbeits- und Ergebnisbericht

Starting Point and Goals of the Project

The goal of the FIBISS project was to develop methods to distinguish false and correct information in social media in the medical domain and to automatically provide potential evidence for this judgement. The main research questions from the computational point of view were: How can we design information extraction models which work both on social media and scientific text? How can we extend information extraction models to accurately estimate the trustworthiness of relations extracted jointly from social media and scientific text? The main goal regarding a practical outcome was to devise a method able to recognize conflicting and unreliable information in the biomedical domain.

Project-specific Results

The structure of biomedical claims has, by and large, not been studied in a focused manner. Therefore, we developed a claim corpus of social media messages in the medical domain [1]. This textual resource consists of 1200 Tweets, annotated with the text span that expresses the main claim. As the domain, we chose various medical conditions.

Our hypothesis in the project was that entities and their relations contribute an important ingredient for the core information claims are made of. We therefore developed an annotated corpus which is, up-to-today, annotated with the largest set of relation categories between entities [2]. This resource of 2100 Tweets with 6,000 entities and 2,200 relations is focused on biomedical claims, but annotated with general biomedical entities. It therefore serves the general purpose of reconstructing patient journeys from the information people share on social media. As a link to fact-checking, we further annotated this corpus with information on the factuality of the claims in the data [3].

To further understand the anatomy of social media health claims and the role of structured knowledge within them, our third published resource contributes the CoVERT dataset [4]. It constitutes the first biomedical fact checking dataset that leverages crowdsourcing to collect annotations and has a focus on COVID19, as an important epidemic event during the project runtime. This resource has been successfully adopted by the community and is established as a standard fact checking benchmark.

To better understand how claims made in social media (and news), the domain in which we perform fact-checking, and the potential evidence (scientific texts), we further performed an annotation and modeling study to understand how journalists or social media channels ‘translate’ a scientific finding to laypeople terms. A reported finding therefore constitutes a type of claim about a finding described in a scientific publication. To detect false claims, it is crucial to understand how accurately the report conveys the original information and across which fine-grained dimensions of information, e.g., causality or certainty, the two may diverge. Therefore, we study which properties within scientific findings are distorted when journalists or social media

users report discoveries [5]. The resulting publication contributes a novel dataset, models and analyses to identify distortions in causality, certainty, generality and sensationalism between two findings, enabling misinformation detection in science communication on a highly fine-grained level.

Based on these resources and modeling experiments with the goal of understanding the phenomenon of claims in social media and the relation to reliable (scientific) resources, we developed a set of methods and integrated systems. We show that a system that uses claim representations based on the entity–relation–entity triple leads to more reliable fact-checking prediction results than using the original formulation in the social media post [6]. In a follow-up study [7], we implement a fully automatic pipeline for real-world biomedical fact checking that leverages the claim extraction method we propose in [5]. The experiments show that fact checking performance improves considerably over tasking the model to make a prediction for an unchanged tweet.

In addition to evidence-based fact-checking, we contribute corpora and results on a related concept, namely deception. To understand the relation between deception and fact-checking, we aimed at answering the question if current fact verification models potentially confound non-propositional cues of deception or the implicit knowledge they store from pertaining into their predictions. This may lead to (a) factual claims being more reliably verified compared to non-factual ones and (b) cause models to perform worse for instances in which the evidence is corrupted by a deceptive intention. We study the connection between fact verification and deception, finding that particularly smaller models show lower performance for instances with non-factual claims and deceptive evidence documents, indicating that these instances are more difficult to verify [8,9,10].

Changes in the project plan

The original plan of the project consisted of five workpackages. WP1 has been designed to focus on data collection and annotation. This workpackage has been conducted as originally planned. WP2 and WP3 has been developed to focus on named entity recognition, entity linking and relation detection methods as the key element to connect information in social media and scientific texts. In the experiments conducted in this workpackage, we found out that entity linking of concepts mentioned in social media to ontologies developed for medical experts only partially allowed an automatic assessment of the required granularity level in automatic entity linking. Therefore, automatic entity recognition systems, while including entity linking, remains a challenge. Hence, we changed the plans as they have been originally developed based on the development of fact-checking systems in WP4. We recognized that these systems work well with automatically reformulated claims, based on the entity extraction in social media. An entity recognition in scientific text, however, turned out to not be required for a successful linking. Therefore, instead of developing an automatic linking procedure between social media–fact-checking–evidence, we decided to develop reformulation methods of the original claim and leave the fact-checker and the evidence in their original form.

All project-related publications are published at highly ranked, peer-reviewed venues within the area of natural language processing as platinum open access. All resources which resulted from the project are available publicly to the community. The resources are linked to at https://www.uni-bamberg.de/en/nlproc/resources/ as well as from the individual scientific papers. Further, the project page at https://www.uni-bamberg.de/en/nlproc/projects/fibiss/ summarizes the project results.

4 Veröffentliche Projektergebnisse

4.1. Publikationen mit wissenschaftlicher Qualitätssicherung

[1] Amelie Wührl and Roman Klinger. 2021. Claim Detection in Biomedical Twitter Posts. In Proceedings of the 20th Workshop on Biomedical Language Processing, pages 131–142, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.15

[2] Amelie Wührl and Roman Klinger. 2022. Recovering Patient Journeys: A Corpus of Biomedical Entities and Relations on Twitter (BEAR). In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4439–4450, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.472

[3] Amelie Wuehrl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024. What Makes Medical Claims (Un)Verifiable? Analyzing Entity and Relation Properties for Fact Verification. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2046–2058, St. Julian's, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.124/

[4] Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 244–257, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.26/

[5] Amelie Wuehrl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding Fine-grained Distortions in Reports of Scientific Findings. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6175–6191, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.369

[6] Amelie Wührl and Roman Klinger. 2022. Entity-based Claim Representation Improves Fact-Checking of Medical Content in Tweets. In Proceedings of the 9th Workshop on Argument Mining, pages 187–198, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics. https://aclanthology.org/2022.argmining-1.18

[7] Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An Entity-based Claim Extraction Pipeline for Real-world Biomedical Fact-checking. In Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), pages 29–37, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.fever-1.3

[8] Aswathy Velutharambath, Amelie Wührl, and Roman Klinger. 2024. Can Factual Statements Be Deceptive? The DeFaBel Corpus of Belief-based Deception. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2708–2723, Torino, Italia. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.243/

[9] Aswathy Velutharambath, Amelie Wuehrl, and Roman Klinger. 2024. How Entangled is Factuality and Deception in German?. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9538–9554, Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.557/

[10] Aswathy Velutharambath and Roman Klinger. 2023. UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 39–51, Toronto, Canada. Association for Computational Linguistics. https://aclanthology.org/2023.wassa-1.5/

4.2 Weitere Publikationen und öffentlich gemachte Ergebnisse

- Corpus on Science Communication Distortions: https://www.uni-bamberg.de/en/nlproc/resources/sciencecommdistortion/
- DeFaBel Corpus of Belief-based Deception: https://www.ims.uni-stuttgart.de/data/defabel
- BEAR, BEAR-Fact, BioClaim, CoVERT Corpora: https://www.ims.uni-stuttgart.de/en/research/resources/corpora/bioclaim/
- UniDecor: A Unified Deception Corpus for Cross-Corpus Deception Detection: https://www.ims.uni-stuttgart.de/data/unidecor