Detecting and Debunking Fake News and Half-truth: A Survey

Singamsetty Sandeep
Department of CSE
IIT Bombay
sandeepsingamsetty000@gmail.com

Pushpak Bhattacharyya
Department of CSE
IIT Bombay
pushpakbh@gmail.com

Abstract

Fake news and half-truths have existed even before the digital era. However, with the rapid rise in internet usage, social media users, news channels, and digital platforms, the spread of fake news and half-truths has become faster. Fake news can be entirely false, while half-truths are partially true or manipulated to misrepresent the truth. Spreading fake news is the easiest way to gain viewership, engage with users, and advertise digitally. The dissemination of fake news and half-truths carries several downsides as it can disrupt social and economic harmony. This paper presents a comprehensive survey of the past works and datasets that exist in the domain of fact-checking, fake news, and half-truth. This paper serves as a roadmap to explore past works and to further build upon them.

1 Introduction

The dissemination of disinformation, especially in the form of half-truths, can have significant and negative implications as it has the potential to disrupt social and economic harmony (Allcott and Gentzkow, 2017; Su et al., 2020). A recent example of this was seen during the Covid-19 vaccination drive, where the spread of disinformation led to widespread fear and skepticism among the public regarding the efficacy and safety of the vaccine (He et al., 2021; Shahi and Nandini, 2020).

Our work tackles half-truths by utilizing the LIAR-PLUS dataset (Alhindi et al., 2018) for half-truth detection. There are many forms of half-truth such as deception, exaggeration, propaganda, and intentionally hidden facts, etc. In this work, we only deal with half-truths related to deception and intentionally hidden facts. Our approach not only detects half-truths but also aims to debunk the claim by editing and transforming it into a truthful statement. ***'Claim'***, as coined by (Toulmin, 2003), is *'an assertion that deserves our attention'*. In our study, a *claim* is defined as a textual statement that can be made by individuals, news websites, political parties, and other sources.

To combat fake news and half-true news, we must detect them faster. Traditionally, fact-checking is done manually. It is time-consuming. The fact-checkers take a lot of time to find the evidence and validate it. Evidence extraction is necessary to automate the process of evidence collection and collect it faster. In our work, we use real-time evidence extraction using Google news scraper.

The main intention of people who spread fake news is to gain attention to boost the number of hits on their websites increase their views and so on. The downside of spreading fake news is a lot of people get surprised and excited by the news that is being circulated, and people tend to circulate the same news or share content with other people in many WhatsApp groups or social media accounts or on Twitter by retweeting. When this news becomes viral, the spread of fake news is faster than ever. Hence, it is very important to keep a check on the spread of fake news and to debunk this fake news at a very early stage so that less fake news can be circulated further and fake news that has been circulated can be countered with the spread of the debunked true news.

This research is a significant advancement in the field of natural language processing (NLP) and has the potential to contribute to fact-checking and computational journalism, ultimately helping to prevent people from falling prey to disinformation.

1.1 Motivation

The spread of disinformation on digital platforms has become a common tactic to attract more viewership. However, traditional fact-checking methods rely on human fact-checkers and can be time-consuming (Hassan et al., 2015), which limits their effectiveness in responding to the constant stream

of disinformation. This is where automated fact-checking (Guo et al., 2022) and disinformation debunking systems become crucial, as they can quickly detect (Monti et al., 2019) and respond to disinformation in real-time, which can help limit its reach (Cohen et al., 2011). Claim editing is an important step in debunking since we can convert fake news and half-true news into real news using the evidence we collected. The edited claims can then be used to counter fake news and half-true news by publishing an article with the edited claim as the headline.

1.2 Problem Definition

The main is to detect fake news and half-truth faster and debunk them. In our work, we have implemented a disinformation detection model to detect disinformation. Given a claim C and the corresponding evidence E as input, the disinformation detection model predicts whether the given claim is true or half-true, or false. It is a three-class classification problem.
In addition to that, we have implemented a claim editing model to edit *half-true* and *false* claims. Given a *half-true* or *false* claim C and the corresponding evidence E as input, our claim editing pipeline uses the evidence to edit the *half-true* or *false* claim and tries to generate an edited *true* claim $C*$ with control over editing the selected parts of input claim. The overall task is depicted in Figure **??**.

2 Background and Terminology

In this section, we shall cover the background, terminology, and definitions of the important concepts in this research.

2.1 Fake News

Fake news is generally false or misleading information that is presented as if it is true news. Fake news has no basis and in fact, it is presented as being accurate by many fake news websites, and often it is found in social media.

2.1.1 Examples

Example 1: *Ravindra Jadeja is out of the Chennai Super Kings team from the 2023 IPL.*

The above news is completely fake till the date this report is written. This is just false propaganda. This is published just to grab the attention of the users and increase the views of those fake articles.

Example 2: *India's national anthem is recognized as the best national anthem of the world by the United Nations.*

The above news is completely fake till the date this report is written. This was trending in social media in early 2010. This is just false propaganda. This news grabbed a lot of attention and many people shared this news across various platforms. Later this news was found to be false.

2.2 Half-truth

A ***half-truth*** is a statement that is partially true but intentionally omits important details that would significantly alter its meaning. This type of statement is deceptive as it can lead to misunderstandings or false impressions. Even if a statement is technically true, it cannot be considered entirely truthful if it excludes crucial information. Half-truths are lies of omission.

2.2.1 Examples

Here are a few examples of half-true statements.

Example 1:*Electronic gadgets mandatory for e-census in 2023.*

The above news is half-true since it is partly true and uses deception. It is hiding the important information that the gadgets will be made available by government officials and the public need not own them. This information is extremely important, else it might confuse the end-user who is reading this news.

Example 2: *I have never purchased a train ticket in my life to travel.*

The above statement might be completely true but doesn't convey complete information and is misleading. What if this person has never traveled on a train? In that case, there is no need to buy a train ticket. Show the important information that the person has never traveled on a train is being hidden and it is conveyed in a negative sense that this person has traveled in a train without purchasing a ticket. Hence the statement even though is

entirely true is also considered a half-truth because it doesn't convey the complete information.

Example 3: *People in Cuba are stinging themselves with blue Scorpions.*[1]

People in Cuba use an antidote to boost immunity. This antidote is made from the poison of blue Scorpions. But the above statement conveys that the people are directly bitten by Scorpions which is an exaggeration of the original situation. Hence this is also considered a half-truth.

Example 4: *Aswattama Hathaha! (Kunjaraha)*

The above example is from Mahabharatha. Yudhishthira, the elder brother of the Pandavas was forced to lie that Ashwattama, son of Dronacharya, is dead. But Yudhishthira being the follower of Dharma never lied. So, he said Aswattama Hathaha, loudly and Kunjaraha (an elephant) slowly. Here he made a deceptive statement, which is half-true.

2.3 Fact-checking pipeline

A fact-checking pipeline typically consists of the following stages. Please refer to figure 1 to have an idea about the different stages of the fact-checking pipeline.

2.3.1 Claim Detection

The claim detection stage filters all the check-worthy claims, since only claims that are worth checking need to be verified. For simplicity, we have assumed all the claims in the LIAR-PLUS dataset are check-worthy.

2.3.2 Evidence Extraction

The evidence extraction stage extracts evidence for each claim from a trustworthy source. In our case, we used justifications extracted from the **PolitiFact** website for verification of claims of the LIAR-PLUS dataset. Later, we developed a real-time evidence extractor using a Google News scraper.

2.3.3 Claim verification

The claims and corresponding evidence were later validated. This stage is called claim verification. At this stage, we verify the claim based on the evidence we extracted and produce a verdict with a justification for the verdict.

- Verdict prediction: Verdict prediction is the stage where we predict the veracity of the claim. Veracity is the degree of truthfulness. We have used a BERT-based model for veracity prediction.
- Justification production: Just producing the veracity label is not sufficient. Hence we also produced an explanation for the predicted veracity label in the form of supports or counters. We have used an NLI model for justification production which uses the idea of textual entailment.

3 Related Work

This section presents the literature survey. This section covers the work from which our research draws inspiration and also covers the work which is similar to our work against which we compete to make our work better.

3.1 Foundational research

This section covers the foundational research papers from which we have drawn inspiration for our work.

3.1.1 Fact Checking

- The survey conducted by (Guo et al., 2022) provides a comprehensive examination of the models and datasets prevalent in the field of fact-checking. This paper meticulously outlines the various challenges encountered within this domain and also offers insights into potential future directions. The concept of debunking false information has been derived from the future directions section of this survey. The challenges enumerated in this study serve as a valuable resource, presenting a holistic overview of the intricate problems that could be addressed in forthcoming research endeavors.
- The survey conducted by (Kotonya and Toni, 2020) focuses on several techniques employed in explaining the verdicts generated by automated fact-checking systems. This paper serves as a source of inspiration for providing explanations in the form of counters and

[1] https://www.reuters.com/article/us-health-cancer-cuba-scorpion-idUSKBN1OD2GH

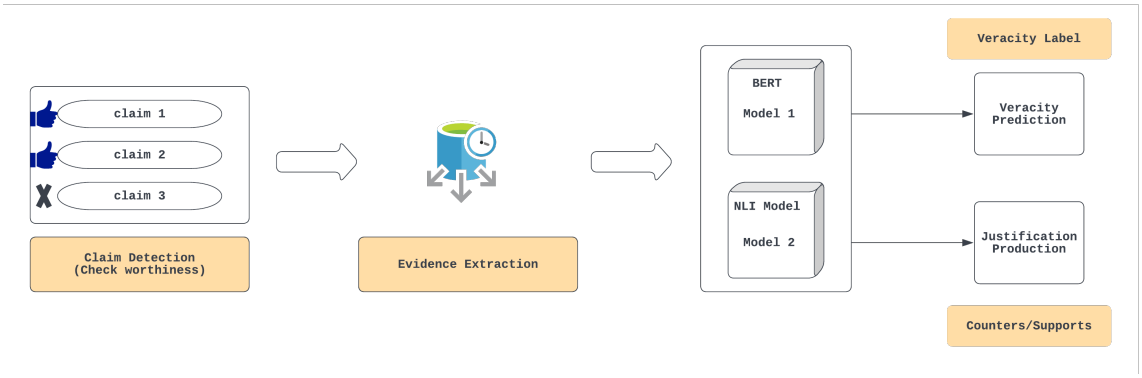

Figure 1: Fact-checking pipeline

supports for fake news. Although the specific idea is not explicitly mentioned, the paper presents various works that utilize different mechanisms to offer explanations. It comprehensively covers almost all the techniques existing in the fact-checking domain for explaining verdicts.

3.1.2 Explainable Fact Checking

- The paper (Atanasova et al., 2020a) proposes a method for generating justifications for claims, where the generated textual summary serves as an explanation for the predicted veracity label of the claim. The notion that a textual summary or sentence can be employed as an explanation was derived from this paper and applied judiciously in our own research.

- The paper (Gardner et al., 2020) demonstrates the efficacy of contrast sets by generating them for diverse datasets. This concept has been utilized to explore counterfactuals, and subsequently, the idea of debunking fake news using counterfactuals emerged. Thus, this paper proved valuable in offering ideas and enhancing our understanding in this area.

- The paper (Atanasova et al., 2020b) presents a technique for generating adversarial examples by utilizing an extended version of the HotFlip algorithm to target the label of each claim in the FEVER dataset. The inspiration for modifying structural components with minimal edits while preserving the content to create edited claims was derived from this research. Although our method differs, the fundamental principles and ideas have been drawn from this paper.

3.1.3 Checkworthiness

The paper (Wright and Augenstein, 2020) focuses on detecting check-worthiness for claims in multiple datasets and demonstrates superior performance compared to benchmark models.

3.1.4 Half-truth

Estornell et al. (2019) discusses the computational complexity of deception by half-truth. The authors demonstrated that half-truths can be computationally more challenging to detect than other forms of deception, thus emphasizing the need for specialized approaches to identify and address this issue. Building on this idea, Monteiro et al. (2018) filtered out half-truths during fake news detection and expressed their idea of detecting half-truths in the future. Motivated by this idea, our work attempts to address the half-truth detection problem.

Along with half-truths, there are other forms of disinformation, such as fake news and exaggerated and sensationalized news. Wright and Augenstein (2021a) focuses on detecting exaggeration in the claims made by press releases. The authors propose a supervised learning approach that utilizes sentence-level features to detect exaggerated claims. Li et al. (2017) conducted an analysis and inspection of exaggerated claims in the domain of scientific news. The authors proposed a framework that leverages natural language processing techniques to detect exaggerated claims in scientific news articles.

3.2 Competitive research

This section covers the competitive research papers that challenge and motivate us to produce better results.

3.2.1 Interpretable fact-checking and claim editing

- (Chi and Liao, 2022) discusses a few ideas about interpreting the predicted label and they use dialog trees to achieve this. This is mostly used for social media data. This paper is a competitor for our research since we also focus on providing explanations for claims which are fake from social media. The idea is certainly different from our idea but it is a sophisticated idea with a good mathematical formulation. It also uses a tree kind of structure to give explanations with the added advantage of using metadata of social media very smartly.

- (Ross et al., 2021) is a semantically controlled text generation system that uses SRL tags smartly and creates contrast sets for various downstream tasks without separately training a model for each task. We have used this idea from the tailor and developed a stronger system than the tailor using a paraphrase dataset to train our model. But Tailor is certainly a competitor because of the additional functionalities that it has got along with maximizing the context to edit claims.

3.2.2 Veracity Prediction

- The paper (Alhindi et al., 2018) introduces the LIAR-PLUS dataset, which is relevant to our work on veracity detection. We have developed a system that competes with the methodology presented in this paper. Notably, our system achieves higher accuracy compared to the LIAR-PLUS dataset paper.

- The paper (Guo et al., 2019) introduces a significant concept of leveraging emotion for fake news detection. The authors utilize both the emotion of the publisher and social emotion, extracting dual emotion features to enhance existing techniques in fake news detection. This idea of incorporating emotion-based approaches holds promise for our future exploration and investigation.

- The paper (Martinez-Rico et al., 2021) explores several models and techniques aimed at estimating checkworthiness and detecting fake news. Wright and Augenstein (2021a) focuses on the detection of exaggeration in the claims made by press releases in comparison to the scientific claims. Li et al. (2017) has conducted an analysis and inspection of exaggerated claims in the domain of scientific news. This paper performs better in comparison to many baselines.

Split	Count
Train	10240
Test	1283
Validation	1284

Table 1: LIAR-PLUS dataset composition

.

4 Datasets

This section lists the datasets that have been created and used in the domain of fake news, rumor detection, disinformation detection, etc. In our research, we have used a few of these datasets for the evaluation of many tasks and models that we have created.

4.1 LIAR-PLUS

LIAR-PLUS[2] dataset is an extended version of the LIAR dataset. This dataset is introduced in the paper titled **Where is Your Evidence: Improving Fact-checking by Justification Modeling** by (Alhindi et al., 2018). The column, **extracted justification** is the new addition made to the LIAR dataset. The dataset composition is listed in the table 1.

4.2 FEVER

FEVER (Fact Extraction and VERification)[3] by (Thorne et al., 2018) is a collection of more than 185,000 claims generated by modifying sentences collected from Wikipedia and then validated without knowledge of the sentences from which they were derived. The claims are categorized as *Supported*, *Refuted*, or *NotEnoughInformation*.

4.3 FaVIQ

FaVIQ (Fact Verification from Information-seeking Questions)[4] by (Park et al., 2022) is a collection of about 26000 claims and corresponding positive and negative evidence list. The dataset composition is listed in the table 2.

[2] https://github.com/Tariq60/LIAR-PLUS
[3] https://huggingface.co/datasets/fever
[4] https://github.com/faviq

Split	Count
Train	17008
Test	4688
Validation	4260

Table 2: FAVIQ dataset composition

.

Split	Count
Test	11809
Validation	10436

Table 3: DialFact dataset composition

.

4.4 DialFact

DialFact by (Gupta et al., 2021) is a benchmark dataset for fact-checking in dialogue. This dataset contains crowd-annotated conversational claims paired with Wikipedia evidence. The dataset composition is listed in the table 3.

4.5 MT-PET

MT-PET by (Wright and Augenstein, 2021b) is a multi-task version of Pattern Exploiting Training (PET), which is a scientific exaggeration detection dataset. This dataset is studied to understand the role of exaggeration in scientific claims and how deception is used along with exaggeration.

4.6 RumourEval

RumorEval 2017 by (Gorrell et al., 2019) is a dataset of controversial posts on social media and the subsequent dialogues, annotated for both stance and veracity. The breaking news stories that give rise to social media rumors are diverse, and the dataset has been extended to also include Reddit and recent Twitter posts.

4.7 CheckThat!

CheckThat 2020 by (Barrón-Cedeño et al., 2020) is a fact-checking dataset created by claims extracted from social media platforms and news articles from various sources. This dataset is used to check the worthiness of claims.

4.8 COVID-19 Disinformation dataset

COVID-19 Disinformation dataset by (Alam et al., 2021) is a dataset created by extracting tweets about Covid-19 from Twitter and annotated manually for the correctness of the claims and disinformation.

4.9 TAPACO

TAPACO dataset by (Scherrer, 2020) is a free paraphrase corpus for 73 languages extracted from the Tatoeba database. Tatoeba is primarily a crowdsourcing project for language learners.

5 Summary

The introduction chapter of our survey paper delved into the fundamental aspects of the topic, establishing a solid foundation for subsequent discussions. We meticulously defined key terms and introduced essential terminology to ensure a comprehensive understanding of the subject matter. Additionally, we provided an overview of the datasets that were utilized in our research, highlighting their significance in the context of our study.

A significant portion of the chapter was dedicated to examining previous works in the field of fact-checking, fake news, and half-truths. We conducted an extensive review of existing literature, meticulously examining and presenting a thorough coverage of prior research conducted in this domain. This comprehensive analysis enabled us to contextualize our own work and identify gaps or areas of further exploration.

Overall, our survey paper's introduction chapter effectively laid the groundwork for the subsequent chapters, ensuring that readers gain a clear understanding of the terminology, datasets used, and the existing body of knowledge related to fact-checking, fake news, and half-truths.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP 2021*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. *CoRR*, abs/2004.05773.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat 2020: Automatic identification and verification of claims in social media. *CoRR*, abs/2007.07997.

Haixiao Chi and Beishui Liao. 2022. A quantitative argumentation-based automated explainable decision system for fake news detection on social media. *Knowledge-Based Systems*, 242:108378.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational journalism: A call to arms to database researchers. In *CIDR*.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2019. Deception through half-truths. *CoRR*, abs/1911.05885.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating NLP models via contrast sets. *CoRR*, abs/2004.02709.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media. *CoRR*, abs/1903.01728.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. *CoRR*, abs/2011.03870.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Juan R. Martinez-Rico, Juan Martínez-Romo, and Lourdes Araujo. 2021. Nlp&ir@uned at checkthat! 2021: Check-worthiness estimation and fake news detection using transformer models. In *CLEF*.

Rafael Monteiro, Roney Santos, Thiago Pardo, Tiago Almeida, Evandro Ruiz, and Oto Vale. 2018. *Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, pages 324–334.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *CoRR*, abs/1902.06673.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: Fact verification from information seeking questions. In *ACL*.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *CoRR*, abs/2107.07150.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid - A multilingual cross-domain fact check news dataset for COVID-19. *CoRR*, abs/2006.11343.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research*, 1:1–13.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Dustin Wright and Isabelle Augenstein. 2020. Fact check-worthiness detection as positive unlabelled learning. *CoRR*, abs/2003.02736.

Dustin Wright and Isabelle Augenstein. 2021a. Semi-supervised exaggeration detection of health science press releases. *CoRR*, abs/2108.13493.

Dustin Wright and Isabelle Augenstein. 2021b. Semi-Supervised Exaggeration Detection of Health Science Press Releases. In *Proceedings of EMNLP*. Association for Computational Linguistics.