Computational Models of Argument Structure and Argument Quality for Understanding Misinformation

Tariq Alhindi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Computational Models of Argument Structure and Argument Quality

for Understanding Misinformation

Tariq Alhindi

With the continuing spread of misinformation and disinformation online, it is of increasing importance to develop combating mechanisms at scale in the form of automated systems that can find checkworthy information, detect fallacious argumentation of online content, retrieve relevant evidence from authoritative sources and analyze the veracity of claims given the retrieved evidence. The robustness and applicability of these systems depend on the availability of annotated resources to train machine learning models in a supervised fashion, as well as machine learning models that capture patterns beyond domain-specific lexical clues or genre-specific stylistic insights. In this thesis, we investigate the role of models for argument structure and argument quality in improving tasks relevant to fact-checking and furthering our understanding of misinformation and disinformation. We contribute to argumentation mining, misinformation detection, and fact-checking by releasing multiple annotated datasets, developing unified models across datasets and task formulations, and analyzing the vulnerabilities of such models in adversarial settings. We start by studying the argument structure's role in two downstream tasks related to fact-checking. As it is essential to differentiate factual knowledge from opinionated text, we develop a model for detecting the type of news articles (factual or opinionated) using highly transferable argumentation-based features. We also show the potential of argumentation features to predict the

checkworthiness of information in news articles and provide the first multi-layer annotated corpus for argumentation and fact-checking.

We then study qualitative aspects of arguments through models for fallacy recognition. To understand the reasoning behind checkworthiness and the relation of argumentative fallacies to fake content, we develop an annotation scheme of fallacies in fact-checked content and investigate avenues for automating the detection of such fallacies considering single- and multi-dataset training. Using instruction-based prompting, we introduce a unified model for recognizing twenty-eight fallacies across five fallacy datasets. We also use this model to explain the checkworthiness of statements in two domains.

Next, we show our models for end-to-end fact-checking of statements that include finding the relevant evidence document and sentence from a collection of documents and then predicting the veracity of the given statements using the retrieved evidence. We also analyze the robustness of end-to-end fact extraction and verification by generating adversarial statements and addressing areas for improvements for models under adversarial attacks. Finally, we show that evidence-based verification is essential for fine-grained claim verification by modeling the human-provided justifications with the gold veracity labels.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I want to express my deepest gratitude to my advisor, Smaranda Muresan, for her support, patience, and guidance. Smara has been the best mentor I could ask for. She was always available to meet and discuss research directions, and she would immensely help in articulating research ideas into clear project proposals. She has encouraged and enabled collaborations with people within and outside of Columbia, which was crucial to my research. Her comments and edits on research papers are magical touches that transform wordy and unclear writing into concise and inspiring work. I am a much better and more mature researcher because of what I learned from Smara.

I would also like to thank the rest of my thesis committee members Preslav Nakov, Daniel Preoţiuc-Pietro, Kathleen Mckeown, and Julia Hirschberg. I had the privilege to work on a research project with Preslav, a pioneer in fake news and misinformation detection, and I learned a lot from him in that research area. Daniel joined Smara and me on an ongoing project and made the work much more solid and generalizable with his valuable contributions and remarks. I also thank Kathy and Julia for their constructive feedback and guidance.

I extend my appreciation to King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia, for providing me with a scholarship that covered my graduate studies for five years.

Throughout my Ph.D., I have been fortunate to work with many people at Columbia and other institutions. Special thanks to my co-authors and collaborators: Elena Musi, Tuhin Chakrabarty, Christopher Hidey, Debanjan Ghosh, Amal Alabdulkarim, Muhammad Abdul-Mageed, Savvas Petridis, Brennan Xavier McManus, Kriste Krstovski, Ali Alshehri, and Jonas Pfeiffer. Thank you for your ideas and work that helped foster my research. I have learned a lot from all of you.

I would also like to thank the many talented colleagues I met at Columbia: Ramy Eskander, Olivia Winn, Giannis Karamanolakis, Sakhar Alkhereyf, Noura Farra, Chris Kedzie, Elsbeth Turcan, Fei-Tzin Lee, Emily Allaway, Katy Gero, Oscar Chang, Lampros Flokas, Brian Goodchild and Tom Effland. Thank you for your thoughts and perspective during our discussions.

Thank you to all of my NYC friends, especially: Saleh Alghusson, Abdullah Altorbaq, Mashal Alrajhi, Sultan Albeshri, Hani Altwaijry, Omar Aldabbah, Omar Alhazzaa, Ahmed Almadan, and Omar Sagga. You have made my time in the city more memorable and enjoyable.

My final and most important thanks go to my family: my parents, my siblings, my wife, and my son. My father, Abdullah, has always been the biggest champion and supporter of higher education. My mother, Alanood, is the warm corner and refuge of my life even as I got older. My siblings, Ali, Hadeel, and Faisal, are my special friends and buddies who have always been there for me. My 2-year-old son, Abdullah jr. (aka Aboodi), has been the highlight and the biggest joy of the past two years. My wife, my partner in life, and my best friend, Batool, thank you for your patience with the long work on some nights, your inspiration and support during the Ph.D., and your warm company and presence in my life. None of this would have been possible without you.

To my parents, siblings, son, and my wife Batool.

Chapter 1

Introduction

In the years 2016 and 2017, the words "Post-Truth" and "Fake News" were named words of the year by the Oxford[1] and Collins[2] dictionaries, respectively. In 2020, paired with the urgent calls by the World Health Organization (WHO) to fight the Covid-19 pandemic, WHO declared with similar urgency the need to fight the "infodemic", referring to the (mis-/dis-)information spreading around Covid-19 at that time (Ghebreyesus, 2020). The consequences of false and misleading information are significant, including swaying political elections (Wardle, 2016), undermining issues such as climate change (Adl-Tabatabai, 2016), and even hospitalization during the early stages of the Covid-19 pandemic (Islam et al., 2020). Thus, with the amount and pace of information spread online, it is important to develop computational models to help detect fake news, understand its nature, and ultimately help increase the digital literacy of users online to reduce their vulnerability.

The term "Fake News" is a vague umbrella to describe false, fabricated, and misleading news, which became a weaponized term to discredit journalists rather than describing the content (Ireton and Posetti, 2018). It is, therefore, more appropriate to use more precise terms such as misinformation *"misleading information with no intent to harm"*, disinformation *"false information with the intent to harm"*, and malinformation *"true information with the intent to harm"* (Ireton and Posetti, 2018; Wardle, 2020) as shown in Figure 1.1.

In this thesis, we study the linguistic content of (mis-/dis-)information, focusing on modeling argument structure and quality for the task of detecting (mis-/dis-)information. Intent, malinformation (e.g., harassment), and multimodal fake content that includes images and videos are beyond the scope of this thesis. We also do not consider the time a certain claim was made to determine its

[1] https://languages.oup.com/word-of-the-year/2016/

[2] https://web.archive.org/web/20171102214325/https://www.collinsdictionary.com/woty

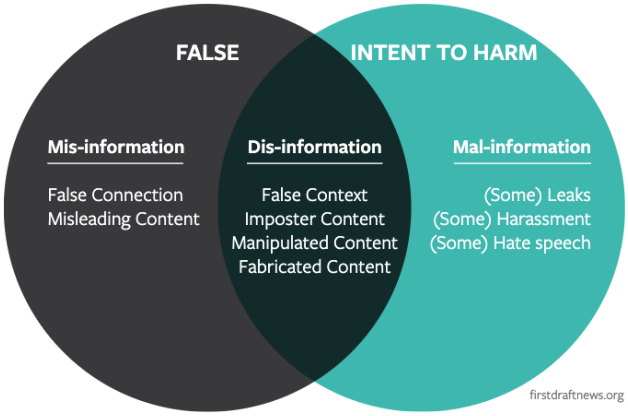


Figure 1.1: The (Mis-/Dis-/Mal-)information space (Ireton and Posetti, 2018).

veracity. To address the temporal effect on the veracity of information, we retrieve evidence from the same timeframe of the claim when possible and leave more in-depth handling for future work.

We study the detection of (mis-/dis-)information with a general objective of facilitating the automation of the fact-checking pipeline, which typically consists of the following four tasks: i) finding statements to fact-check (checkworthiness prediction which assumes both what to fact-check and why); ii) deciding whether they have been previously checked (verified claim retrieval), and if not; iii) retrieving evidence relevant to the target statements; iv) assigning a veracity label to the target statement given the retrieved evidence (claim verification) (Barrón-Cedeno et al., 2020). In this thesis, we particularly focus on the role of argument structure and argument quality in improving the first task and the last task in the fact-checking pipeline.

Consider the example shown in Figure 1.2, where we show part of a news article about climate change with three layers of interconnected annotations. The first layer (red) displays a sample of the argument structure in the article with one claim, two premises, one support relation, and one attack relation. The second layer (blue) highlights three fallacious segments of the article that contain fallacies such as Red Herring *"presenting irrelevant information"*, Strawman *"misinterpreting arguments by others"*, and Loaded Language *"influencing through phrases with strong emotional implications"*. The third layer (green) shows two checkworthy statements with Wikipedia as a potential source of evidence to investigate the veracity of these statements. We study all three layers in this thesis, considering their relation and overlap.

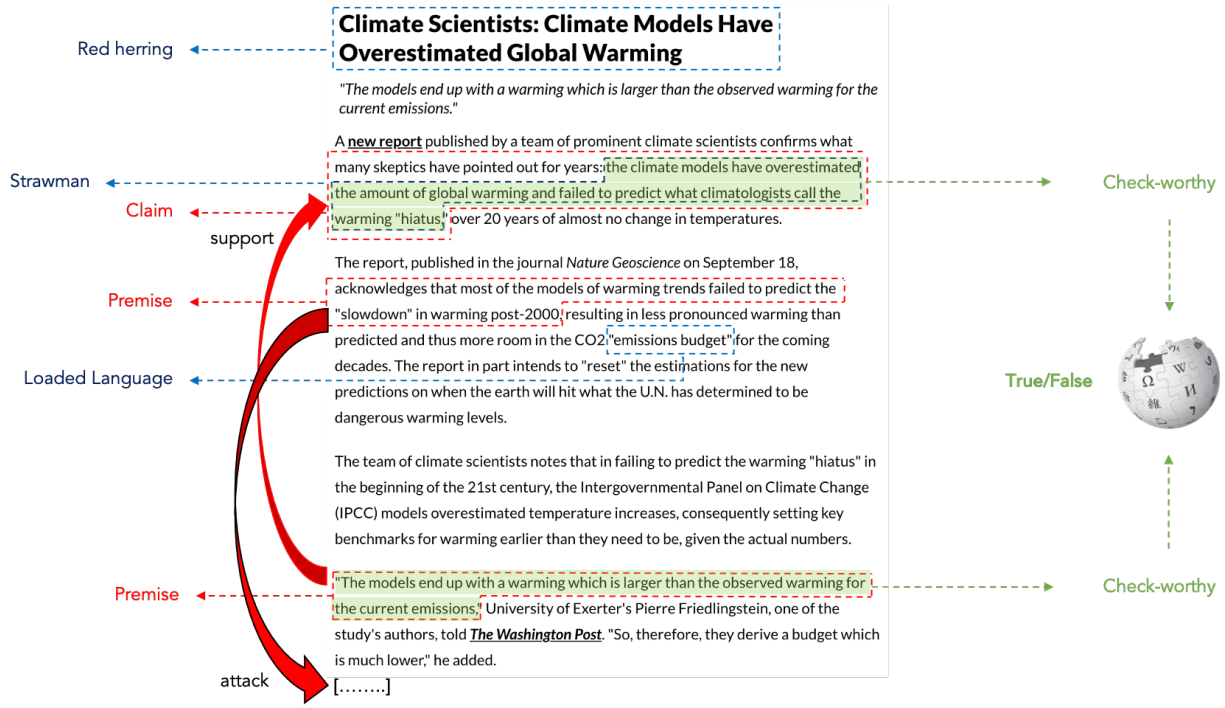


Figure 1.2: A news article with three layers of tagged fragments: Green (checkworthy statements fact-checked against Wikipedia), Red (sample from the argument structure with a claim, premises, a support relation, and an attack relation), and Blue (three types of fallacies).

With the recent progress in deep learning models for Natural Language Processing (NLP), including large language models (e.g., (Peters et al., 2018; Devlin et al., 2019a; Raffel et al., 2020; Brown et al., 2020)), we are interested in complementing the knowledge captured by these models with additional sources in the form of argumentative discourse structure, and instruction-based prompts. We study a wide range of tasks related to misinformation detection and fact-checking, such as: distinguishing factual statements from opinions, assessing the checkworthiness of information in news articles, developing a unified model based on multitask instruction-based prompting for fallacy recognition, and developing claim verification approaches given automatically retrieved (or provided) evidence under truth barometers with different levels. In particular, we address the following research questions:

- Given argumentative discourse structures (through gold annotations or model predictions), how can we utilize them in tandem with deep learning models for NLP to improve on tasks relevant to fact-checking such as separating news from opinions and determining the

checkworthiness of information?

- Can we develop a unified model for fallacy recognition considering variations in the type and number of fallacy classes across datasets covering multiple domains and genres?
- What sources of knowledge (fallacy, human justification, automatically retrieved evidence) are most useful for claim verification models under different truth barometers? And how resilient are these models under adversarial attacks?

The following three sections describe our approaches to answering the three research questions above, respectively.

1.1 The Role of Argument Structure in Fact-Checking

To develop methods for detecting (mis-/dis-)information that contribute to automating the fact-checking process, we need to be able to separate factual statements from opinions. In addition, having a set of factual claims poses the need to determine which of them must be selected for fact-checking due to their "checkworthiness", which indicates factual claims that should be checked to see if they are true. The notion of checkworthiness varies greatly depending on the context (Wright and Augenstein, 2021), however according to Shaar et al. (2020), the following five elements make a statement checkworthy: i) contains a verifiable factual claim; ii) likely to be false; iii) of interest to the general public; iv) harmful to society; v) worth manual fact-checking.

Towards achieving these two goals: separating facts from opinions, and determining the checkworthiness of statements, we investigate the role of argumentative discourse structures. An argument consists of argumentative components (e.g., claims, premises) that are connected through relations (e.g., support, attack) forming a tree or a graph structure (Stede and Schneider, 2018; Lawrence and Reed, 2020). The ability to mine the argument structure improves on several tasks, such as essay scoring and writing assistance (Wachsmuth et al., 2016; Zhang and Litman, 2020). We hypothesize the relevance of the argument structure to the two aforementioned tasks.

First, as opinionated articles have the ultimate goal of persuasion, we expect their argumentative structure to differ from factual articles that focus mainly on reporting. Therefore, we hypothesize that features extracted from the argument structure help separate facts from opinions. Second, the five elements for checkworthiness mentioned above consider only the target statement with no regard to its context. When we look at the context around the manually fact-checked segments by experts in news articles about climate change, we notice patterns related to the argumentative structure. For example, human fact-checkers tend to fact-check a claim when it is not supported by evidence (premise) or only supported by another claim (thus showcasing an "*evading the burden of proof*" fallacy), and fact-check a premise when it is used to support a claim (e.g., to challenge the relevance of that evidence in support for the claim).

Chapter 3 shows the role of argument structure in improving these two tasks.

1.2 Fallacies as Indicators of Misinformation

After considering the role of the structure of the argument in improving tasks relevant to fact-checking, we focus next on qualitative aspects of arguments through fallacies.

A fallacy is *"an argument that seems valid but is not"* used to support a position and persuade an audience about its validity, similar to valid arguments (Hamblin, 1970). It is *"a particular kind of egregious error, one that seriously undermines the power of reason in an argument by diverting it or screening it in some way. But a more precise definition is difficult to give and depends on a range of considerations"* (Tindale, 2007). However, a fallacy *"must have the appearance of being a better argument than it really is"* (Hansen, 2002), and thus it is not easy to detect for both humans and machines. Fallacies apply to various aspects of argumentation: they can apply to the structure (e.g., lack of arguments in support of a claim); to the reasoning linking the argument to the conclusion (e.g., presence of a false cause-effect relation); to language features (e.g., use of vague terms) as well as factors which lie outside discourse (e.g., personal qualities of the protagonists to the discussion) (Tindale, 2007).

Brennen et al. (2020) show that 59% of fact-checked news are not simply true or false, but ones with misleading content or false context thus containing fallacious reasoning. For example, a Cherry-Picking fallacy *"presenting information that supports a position and ignoring others that do not"* is based on presenting partial information rather than false information (Musi et al., 2022). Thus, we hypothesize the relevance of fallacies to fact-checking tasks such as checkworthiness prediction and claim verification as fallacies can be used as indicators of misinformation.

We collaborate with Musi et al. (2022) to develop an annotation scheme of fallacious moves in misinformation and use it to annotate two datasets in climate change and Covid-19 that we present in Chapter 4. In addition, we present a unified model for fallacy recognition through multitask instruction-based prompting across five datasets considering four different formulations of fallacies. We finish the chapter by studying the use of fallacies as indicators of checkworthiness, to provide the reasoning behind why a certain statement should be fact-checked.

1.3 Verification of Statements

In Chapter 5, we present our work for automating the evidence extraction and claim verification process. We study the case when a model is given a claim to fact-check and tasked with predicting the veracity of the claim under different levels of truth, and availability of evidence. First, we present our models for end-to-end fact-checking where a model needs to find relevant evidence from a knowledge base of facts (e.g., Wikipedia), reason about what can be inferred from the evidence with respect to the validity of the claim (Section 5.1). Then, we present a set of adversarial attacks by adding alterations to claims from Wikipedia to resemble naturally occurring claims and assess the vulnerability of fact-checking systems (Section 5.2). We end our discussion of verification by developing models for fine-grained claim verification on a six-level truth barometer (e.g., 'mostly-false', 'half-true') that better assess the veracity of misinformation (i.e., misleading statements) and show the importance of evidence for fine-grained claim verification (Section 5.3).

1.4 Thesis Contributions

- We approach fact-checking with a holistic view by developing models for checkworthiness (what to fact-check), fallacy recognition (why to fact-check), and veracity prediction (how to fact-check), in addition to analyzing the relations between these tasks.
- We establish a connection between misinformation and argumentative fallacies by introducing an annotation scheme, datasets, and models for predicting fallacy types of fact-checked content and thus consider fallacies as indicators of misinformation.
- We combine language models with contextual information to have more robust models on multiple tasks such as:
  - Argument structure for checkworthiness prediction (Alhindi et al., 2021), and news articles type prediction (Alhindi et al., 2020).
  - A multitask instruction-based prompting framework for fallacy recognition across domains, genres, and annotation schemes (Alhindi et al., 2022).
  - Fallacy as a rationale for checkworthiness.
- We release a number of new datasets:
  - A multi-layer annotated corpus for checkworthiness and argumentative discourse structures for climate change news articles (Alhindi et al., 2021), and a fallacy corpus of climate change and Covid-19 news articles and social media posts (Alhindi et al., 2022).
  - The LIAR-PLUS dataset (Alhindi et al., 2018) of fact-checked claims with justifications that is used by the community as a benchmark dataset for developing fact-checking models with generated explanation (Atanasova et al., 2020; Stammbach and Ash, 2020).

1.5 Publications and Thesis Organization

The thesis consists of the following chapters: In Chapter 2, we present a literature review of argumentation and fallacy mining, fake news and misinformation detection, and various attempts to automate part or all of the fact-checking pipeline. Chapter 3 discusses the role of features from argument structure in improving two fact-checking tasks in news articles: separating facts from opinions, and checkworthiness of statements. In Chapter 4, we introduce our study of fallacy and our unified models based on multitask instruction-based prompting for fallacy recognition. Then, we show our work in building end-to-end fact-checking systems under different truth barometers (binary and six-way) in Chapter 5. We end with concluding remarks and future work in Chapter 6. In Table 1.1, we show our contributions and publications for each of the three main chapters.

Chapter		Section	Contribution	Publication
3	Argument Structure in News	3.1	Argumentation features for distinguishing facts from opinion in news.	Alhindi et al. (2020)
		3.2	Multilayer corpus of argument structure and fact-checked content, and argumentation context for predicting checkworthiness	Alhindi et al. (2021)
4	Argument Quality through Fallacy	4.1	A fallacy scheme of fact-checked statements	–
		4.2	Tagging and classification of fallacious segments in a single dataset	Alhindi et al. (2019)
		4.3	Multitask instruction-based prompting for classification of fallacy types across five datasets	Alhindi et al. (2022)
		4.4	Fallacy for explaining checkworthiness of statements	TBD
5	Verification of Statements	5.1	Fact-checking (binary truth labels, or *NotEnoughInfo*)	Chakrabarty et al. (2018)
		5.2	Adversarial attacks against fact-checking systems	Hidey et al. (2020)
		5.3	Predicting shades of truth through human justifications	Alhindi et al. (2018)

Table 1.1: Contributions and publications for each chapter.

It is important to note that the work described in this thesis was done over the span of five years, from 2018 to 2022, in which immense progress has happened in the fields of NLP and machine learning in terms of the approaches used to tackle a new problem and the available models that offer different opportunities for transfer learning. Table 1.1 shows the year when each section of the thesis was done, where the most successful method and language model of that year was used. The thesis is not ordered chronologically but rather in terms of the topic, task, and connection to other parts of the thesis. For example, our earliest work is the one on the verification of statements (Sections 5.1 and 5.3). However, it is presented at the end of the thesis after showing more recent work in the preceding two chapters.

Chapter 2

Related Work

False and misleading content has been classified into different sub-categories in the literature from the three types of fake: *serious fabrication, hoaxes, and satire* (Rubin et al., 2015) to the seven types of (mis-/dis-)information: *satire, misleading content, imposter content, fabricated content, false connection, false context, and manipulated content* (Wardle, 2017). The differences between these categories depend on many factors such as genre and domain, targeted audience, and deceptive intent (Rubin et al., 2015; Rashkin et al., 2017a). A study by Vosoughi et al. (2018) found that false and misleading news spread six times faster than truthful ones on social media. To fight this spread, a plethora of manual fact-checking organizations have emerged or increased their scale in recent years such as PolitiFact[1], FactCheck.org[2], Snopes[3], and FullFact[4]. However, given the amount and scale of information spread online, a need to automate parts or all of the fact-checking pipeline became necessary. To automate fact-checking, Vlachos and Riedel (2014) mapped the process to three natural language processing (NLP) tasks: identifying claims to be checked, finding appropriate evidence, and producing verdicts, while Barrón-Cedeno et al. (2020) adds a fourth task of identifying previously checked claims as false claims tend to be repeated online (Nakov et al., 2021a). More recent work adds a fifth task of producing justifications for the verdict to increase the trust in the predictions of machine learning models (Kotonya and Toni, 2020; Guo et al., 2022). An additional task of claim spotting has been argued recently by Reddy et al. (2022), which considers additional attributes related to checkworthiness such as identifying the claimer, the source of the claim, the stance of the claim, and the claim object.

[1] http://www.politifact.com
[2] http://www.factcheck.org
[3] http://www.snopes.com/
[4] http://fullfact.org

Other work focuses on estimating the credibility of sources by using an external list of bias per publisher (Baly et al., 2018a) or by modeling conflicting reports on a claim from different sources (Zhang et al., 2019b). In addition, there is work that studies other phenomena related to credibility and trustworthiness such as propaganda (Martino et al., 2020), hyperpartisanship (Potthast et al., 2018; Alabdulkarim and Alhindi, 2019), rumors (Zubiaga et al., 2018), and stance detection for (mis-/dis-)information (Hardalov et al., 2022).

Another angle to look at the information ecosystem is through Fallacy Theory. Musi and Reed (2022) considers fallacies as indicators of misinformation and defines a taxonomy of fallacies about misinformation through analysis of 220 news fact-checked by *Snopes* about Covid-19. We collaborate with Musi et al. (2022) to refine and apply this taxonomy on a bigger set of news around Covid-19 from multiple fact-checkers, and on fact-checked news articles about climate change (Alhindi et al., 2022). In addition, Goffredo et al. (2022) shows that looking at the argument structure helps in detecting fallacies in political debates. We investigate the role of using argument structure for detecting structural fallacies (e.g., *Evading the Burden of Proof*), and in facilitating tasks related to fact-checking such as checkworthiness prediction (Alhindi et al., 2021).

In the remainder of this Chapter, we review the literature on the fact-checking tasks covered in this thesis starting with identifying claims to be checked (checkworthiness prediction), and end-to-end fact-checking that includes evidence retrieval and claim verification. Then we briefly introduce argumentation mining (extracting argument structure) and its applications in general and to tasks that can inform fact-checking. Finally, we end with a discussion of argument quality, existing fallacy datasets and previous work on models for fallacy recognition covering different fallacy schemes that include propaganda techniques.

2.1 Checkworthiness Prediction

Previous work on detecting checkworthy claims focuses on text from the political domain. The first systems for checkworthy claim detection are ClaimBuster (Hassan et al., 2017) and ClaimRank (Jaradat et al., 2018). ClaimBuster is trained on sentences from political debates and uses sentence-

level features such as TF-IDF weights and sentiment. ClaimRank extends this to Arabic (in addition to English) and uses a richer feature set that includes the context. Other more recent work includes datasets that are bigger in size and across longer time spans (Arslan et al., 2020) or in other languages such as Dutch (Berendt et al., 2020). Covering multiple domains (political speeches, tweets, Wikipedia) and task formulations (checkworthiness, rumor detection, and citation detection), Wright and Augenstein (2020) use positive unlabelled learning (Bekker and Davis, 2020) to perform a comparison of datasets across domains where the notion of checkworthiness varies greatly.

Over the past five years, the CLEF check-that lab introduced tasks for detecting checkworthy political claims from debates and social media (Nakov et al., 2018; Elsayed et al., 2019; Barrón-Cedeno et al., 2020; Nakov et al., 2021b; Nakov et al., 2022), where the best teams in the 2019 task (Hansen et al., 2019) use syntactic features and word embeddings in an LSTM model. More recently on the same datasets, Kartal et al. (2020) introduce a logistic regression model using BERT-based features, the presence of comparative and superlative adjectives, augmented with data from controversial topics. Finally, Meng et al. (2020) uses adversarial training on transformer neural network models for detecting checkworthy statements. However, all of these models are trained on political text from debates, speeches, and tweets, or lists of claims previously checked by various fact-checking agencies such as `FactCheck.org`. We on the other hand work on a dataset from a different genre: *news articles*, include a new domain: *climate change*, investigate the question of whether argumentative discourse structure helps in detecting checkworthy statements, and study the reasons behind checkworthiness through the analysis of fallacies in checkworthy statements.

2.2 End-to-End Fact-Checking

Starting with a set of claims to fact-check, previous work on predicting the veracity of claims focused on comparing them against evidence from Wikipedia (Thorne et al., 2018a), trusted news outlets (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017a), discussion forums (Joty et al., 2018), or debate websites (Chen et al., 2019), or by analyzing the linguistic properties of false

and true claims (Pérez-Rosas et al., 2018a; Rashkin et al., 2017b) in addition to the speaker's history (Wang, 2017). These datasets are labeled using three tags (*true*, *false*) (Alhindi et al., 2018), three tags (*supported*, *refuted*, *not-enough-information*) (Thorne et al., 2018b), four tags (*agree*, *disagree*, *discuss*, *unrelated*) (Pomerleau and Rao, 2017b), or *Politifact*'s six tags: *pants-on-fire, false, mostly-false, half-true, mostly-true and true* (Rashkin et al., 2017a; Wang, 2017). They vary in size from 300 claims (Ferreira and Vlachos, 2016) to $185,000$ claims (Thorne et al., 2018b).

Claim verification approaches include stylometric and linguistic analysis of the target content, and the comparison of such content against trustworthy evidence (Thorne and Vlachos, 2018; Potthast et al., 2018). We first cover the work on claim verification that analyzes linguistic and stylometric properties of the claims (*stylometry-based verification*). Then, we overview the work that uses evidence to evaluate the veracity of claims (*evidence-based verification*).

2.2.1 Stylometry-based Verification

Several studies analyzed the language used in the claims to assess their veracity. Rashkin et al. (2017a) presented an LSTM model with maximum entropy to predict the truthfulness of claims in news and fact-checked claims from PolitiFact. They found that first-person and second-person pronouns are used more in less reliable news. Subjective language, superlatives, and modal adverbs are used more in fake news. Words used to offer concrete figures, comparatives, money, and numbers appear more in truthful news. Trusted sources are more likely to use assertive words and less likely to use hedging words. Pérez-Rosas et al. (2018b) show that linguistic properties of deception in one domain might be structurally different from those in a second domain. However, we empirically show the importance of evidence-based verification to go beyond language characteristics of claims that might not be generalizable (Alhindi et al., 2018).

2.2.2 Evidence-based Verification

The verification of claims using evidence either performs evidence retrieval at the document or at the sentence level or assumes the availability of evidence and models the relationship between the

claim and the evidence. One of the earlier works was the Fake News Challenge (FNC) (Pomerleau and Rao, 2017b), which is built by randomly matching claim–article pairs from the Emergent dataset (Ferreira and Vlachos, 2016), which itself pairs 300 claims to 2,500 articles. The task is to predict the stance of the article with respect to the claim whether it *agrees*, *disagrees*, *discusses*, or is *unrelated* to the claim. There are several approaches attempting to predict the stance on the FNC dataset using LSTMs, memory networks, and transformers (Hanselowski et al., 2018a; Conforti et al., 2018; Mohtarami et al., 2018; Zhang et al., 2019a; Schiller et al., 2021; Schütz et al., 2021).

This was followed by work that requires evidence retrieval prior to claim verification after the introduction of tasks such as the first and second Fact Extraction and VERification (FEVER) shared tasks (Thorne et al., 2018a; Thorne et al., 2019). Work on end-to-end fact-checking through the FEVER shared task focused on a pipeline approach of retrieving documents, selecting sentences, and then using an entailment module (Malon, 2018; Hanselowski et al., 2018b; Tokala et al., 2019); the winning entry for the first FEVER shared task (Nie et al., 2019a) used three homogeneous neural models. Other work has jointly learned either evidence extraction and question answering (Nishida et al., 2019) or sentence selection and relation prediction (Yin and Roth, 2018; Hidey and Diab, 2018).

2.2.3 Adversarial Attacks Related to Fact-Checking

Language-based adversarial attacks have often involved transformations of the input such as phrase insertion to distract question-answering systems (Jia and Liang, 2017) or to force a model to always make the same prediction (Wallace et al., 2019). Other research has resulted in adversarial methods for paraphrasing with universal replacement rules (Ribeiro et al., 2018) or lexical substitution (Ren et al., 2019). While our strategies include insertion and replacement, we focus specifically on challenges in fact-checking. The task of natural language inference (Bowman et al., 2015; Williams et al., 2018) provides similar challenges: examples for numerical reasoning and lexical inference have been shown to be difficult (Glockner et al., 2018; Nie et al., 2019b) and improved models on these types are likely to be useful for fact-checking. Finally, (Thorne and Vlachos, 2019) provided a

baseline for the FEVER 2.0 shared task with entailment-based perturbations. Other participants generated adversarial claims using implicative phrases such as "not clear" (Kim and Allan, 2019) or GPT-2 (Niewinski et al., 2019). In comparison, we present a set of attacks motivated by realistic, challenging categories and we develop models to address those attacks.

2.3 Models of Argument Structure and Applications

Argumentation mining is a field concerned with finding argument structure from unstructured text that includes argument components (claim, premises) and relations (support, attack) as covered extensively by surveys such as Stede and Schneider (2018; Lawrence and Reed (2020). A standard argumentation mining pipeline includes four tasks: separating argumentative components from non-argumentative ones, classifying the types of argumentative components, extracting relations between argumentative components, and classifying the types of these relations. We do not introduce a new model for argumentation mining; however, we release a new dataset of annotated argument structures in news articles (that also has fact-checking annotations) and we study the use of argument structures in downstream tasks.

There are many argument mining corpora available on text from multiple genres such as student essays (Stab and Gurevych, 2014), short-texts (Peldszus and Stede, 2015), social-media threads (Hidey et al., 2017), and editorials (Al Khatib et al., 2016). Argumentation mining has been used in applications such as legal decision-making (Moens et al., 2007), document summarization (Kirschner et al., 2015), writing assistance (Zhang and Litman, 2016) and essay scoring (Persing and Ng, 2015; Somasundaran et al., 2016), relevance to essay prompts (Persing and Ng, 2014), opinions and their targets (Farra et al., 2015), and argument strength (Persing and Ng, 2015) among others. (Beigman Klebanov et al., 2017) and (Persing and Ng, 2015) analyzed writing of university students and (Stab and Gurevych, 2017a) used data from "essayforum.com", where college entrance examination is the largest forum. However, argumentation applications in news have been limited to analysis of persuasion in editorials (El Baff et al., 2020) and patterns of argumentative strategies

across topics (Al Khatib et al., 2017). We investigate the predictive power of argumentation-based features in the news domain such as the classification of article types (factual vs. opinions) and the prediction of the checkworthiness of statements in news articles.

2.4 Argument Quality and Fallacy

Previous work has empirically studied qualitative aspects of arguments building on theoretical work on argument quality by evaluating aspects such as persuasion and convincingness of arguments. Wachsmuth et al. (2017a) found a high correlation between expert and crowdsourced annotations of argument quality on one argumentation dataset. To enable computational assessment of argument quality, Wachsmuth et al. (2017b) introduced a holistic view of argument quality by introducing measures covering three dimensions: logical (sufficiency, acceptability, and relevance of supporting evidence), rhetorical (clarity, credibility, appropriateness, and emotional appeal), and dialectical (global assessment of the reasonableness of arguments). On the other hand, another angle of assessment of argument quality is through the study of argumentative fallacies.

There are various typologies of fallacies that address informal logic traditions or rules of ideal critical discussion (Hansen, 1996; Van Eemeren et al., 2002; Tindale, 2007; Walton et al., 2008; Damer, 2012). This intersects with propaganda techniques that focus on faulty reasoning and emotional appeals to accomplish persuasion (Miller, 1939a; Jowett and O'Donnell, 2012; Torok, 2015; Weston, 2018). Fallacy datasets include ones that occur in dialogue (Habernal et al., 2017; Sheng et al., 2021), argument sufficiency (Stab and Gurevych, 2017b), name calling on Reddit (Habernal et al., 2018), non-sequitur fallacy in legal text (Nakpih and Santini, 2020), logical fallacies (Jin et al., 2022), fallacies in misinformation (Musi et al., 2022; Musi and Reed, 2022), propaganda techniques in news articles (Da San Martino et al., 2019b) and memes (Dimitrov et al., 2021), and fallacies in political debates (Goffredo et al., 2022).

Fallacy Recognition Models Previous work on fallacy recognition has tackled one dataset at a time such as the structure-aware classifier to detect logical fallacies by Jin et al. (2022). More

relevant to our work on connecting argument structure to fallacy recognition, Goffredo et al. (2022) proposed a transformer-based model architecture that is fine-tuned on argumentation features and showed the importance of detecting argument components and relations for fallacy recognition. Similarly, we look at argument structure for checkworthiness prediction under the hypothesis of its relevance to detecting the *Evading the Burden of Proof* fallacy (Alhindi et al., 2021).

The majority of previous work on fallacy recognition falls under propaganda technique detection. Da San Martino et al. (2019a) introduced a shared task for propaganda detection that consisted of two subtasks: sentence-level classification (binary classification of sentences into propaganda or non-propaganda) and fragment-level classification (finding propaganda segments in news articles and classifying their types into one of eighteen propaganda techniques). The top teams for the sentence classification use ensemble models of neural networks and logistic regression (Gupta et al., 2019) and data upsampling techniques (Tayyar Madabushi et al., 2019), while for fragment classification the top team uses a 20-way word-level classification based on BERT (Yoosuf and Yang, 2019). We participated in this task and were ranked fifth (out of 13) in the fragment classification (Alhindi et al., 2019). We introduced an LSTM-based tagger with relevant dictionary-based features that resulted in having the highest precision model among all teams (more in Section 4.2). On the same dataset, Da San Martino et al. (2019b) introduced a multi-granularity neural network for finding and classifying propaganda fragments. However, due to the complexity of the task, all methods perform lower than 0.25 overall F1 score in fragment-level classification.

In the next iteration of the shared task, the fragment-level classification task was further split into two subtasks: a span identification (SI) task of finding propaganda fragments in news articles, and a propaganda technique (TC) classification task given a propagandistic fragment (Da San Martino et al., 2020). This new task formulation was introduced to reduce the complexity of the simultaneous tagging and classification, and to allow for in-depth exploration of one subtask. The eighteen propaganda techniques became fourteen with some techniques discarded due to small frequency in the data or merged with other ones due to their similarity (e.g., *Whataboutism, Red Herring,* and *Strawman* are merged under one technique). The top teams in the SI task have used methods such

as a heterogeneous multi-layer neural network with part-of-speech (PoS) and named entity (NE) embeddings into an LSTM tagger (Morio et al., 2020), RoBERTa with self-supervision (Jurkiewicz et al., 2020). The top teams in Technique Classification used an SI model to generate propaganda spans that can be used as silver labels for a RoBERTa model (Jurkiewicz et al., 2020) and the use of RoBERTa with post-processing to handle some propaganda techniques such as *Repetition* (Chernyavskiy et al., 2020).

We did not participate in the second iteration of the propaganda detection shared task nor did we tackle any of the other fallacy datasets in a single dataset setup. Working towards a unified approach for fallacy recognition, we tackle five fallacy datasets in a multitask fashion and we present a unified model for fallacy recognition using instruction-based prompts (Alhindi et al., 2022).

Chapter 3

The Role of Argument Structure in Fact-Checking

Argument Mining is the automatic identification and extraction of argument structures that mainly consist of argumentative components (e.g., claim, premise) and their relations (e.g., support, attack). Several approaches have been developed to extract these components and relations from text (Nguyen and Litman, 2016; Eger et al., 2017; Alhindi and Ghosh, 2021). While we do not propose new models for extracting argument structures in this thesis, we focus on utilizing knowledge of argumentative discourse structures into two downstream tasks that can improve fact-checking. Harnessing argumentation mining in the news domain for applications related to fact-checking is still underutilized. It is primarily limited to creating corpora and analytical studies (e.g., argumentation strategies in editorials (Al Khatib et al., 2016)). We show two tasks related to fact-checking that benefit from argumentation features extracted from news articles: i) distinguishing factual from opinionated articles (Alhindi et al., 2020); ii) predicting the checkworthiness of sentences in news articles (Alhindi et al., 2021), as we describe below.

Subjectivity in news reporting has been rising in recent years, especially in online-only publications (Blake and others, 2019). It was estimated that only 41% of publishers label their type of articles (e.g., editorial, review, analysis), and among those who label the types, there is a lack of consistency and clarity (Harris, 2017). A major finding of a 2018 study led by the Media Insight Project showed that most journalists (nearly 80%) think that their news organizations should clearly mark what is news reporting and what is opinion/commentary in order to combat fake news and gain public trust (The-Media-Insight-Project, 2018). Therefore, we develop models for detecting the type of news articles (news story or opinion piece) by introducing argumentation features as described in Section 3.1 (Alhindi et al., 2020). These models can flag content to readers or fact-checkers who

seek to check verifiable factual information, not personal opinions.

The second task is the checkworthiness prediction of sentences in news articles. Most previous attempts at automating fact-checking focus on the verification of claims against automatically- or manually-retrieved evidence from (trusted) sources such as Wikipedia or news articles from credible publishers (Thorne et al., 2018a; Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017a). However, less attention is given to automatically compiling a list of checkworthy statements that can then be inspected and fact-checked by a human fact-checker (or by a fact-checking system). Several previous studies developed datasets and models for identifying checkworthy statements in political news, debates, and social media (Hassan et al., 2017; Jaradat et al., 2018; Arslan et al., 2020; Nakov et al., 2021b; Nakov et al., 2022), while we look at news articles. We utilize argumentation in selecting the proper context for statements to determine their checkworthiness (Section 3.2). In addition, we release the first multi-layer annotated corpus of fact-checked statements and argumentative discourse structures in news articles (Alhindi et al., 2021).

Our contributions in this chapter are the following:

1. We demonstrate that sentence-level argumentation features derived from predictive models are useful in the downstream task of document-level news vs. opinion classification and transfer well to articles from unseen publishers or domains.

2. We introduce a new dataset of 95 climate change news articles with annotations of fact-checked segments and argumentative discourse structure and introduce a model that incorporates information from argumentative discourse structure to predict the checkworthiness of sentences in those articles.

3.1 Fact vs. Opinion in News Articles

Broadly, there are two types of news articles: 1) opinion articles written to present the opinion of the editor or board and aimed to persuade the readers with respect to a particular point of view, and 2) news stories, which aim to report factual news or events. Other less prominent types in the

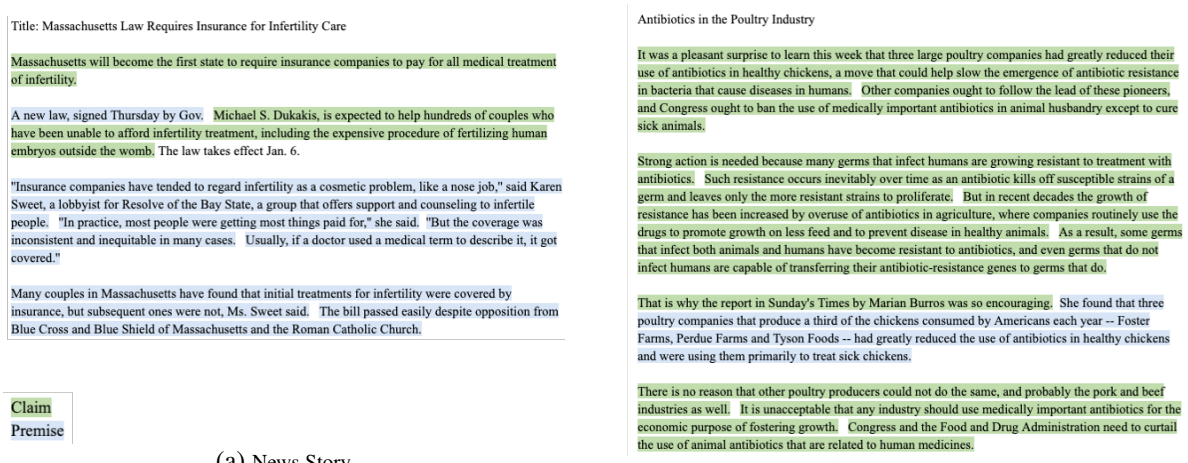
mainstream media such as satire are beyond the scope of this study.[1] Given that the intent of opinion articles is persuasion, we hypothesize that one of the key differences between news stories and opinion articles rests in the discourse structure and, in particular, the argumentative and persuasive aspects of the article. Figure 3.1 shows an example of a news story and an opinion article with two coarse-grained types of argumentative components highlighted (i.e., claims *"stances relating to the text's main issue that needs to be supported"* and premises *"propositions that express reasons to believe a given claim"*). We can see that claims are more prevalent in the opinion article, while the news story contains more premises to support a small number of claims.

We study the predictive power of such coarse-grained argumentation features (claims and premises) for the task of news articles classification into news stories and opinion pieces. For short, we will refer to this binary task as news vs. opinion classification.

To train our sentence-level argument component classification model (claim, premise, none), we use the corpus of editorial news labeled with argumentation strategies introduced by Al Khatib et al. (2016). We compare our approach that uses argumentation features to models using discrete linguistic features from previous work (Krüger et al., 2017) and to document-level transformer-based models such as BERT (Devlin et al., 2019a) fine-tuned for the document-level news vs. opinion classification task. We focus in particular on the transferability of these classifiers, as this task is particularly sensitive to changes in topic or publishers. Therefore, we train and test our models under two regimes. First, we train on articles from one publisher and test on articles from another publisher (including two different domains). For this, we use the dataset introduced by Krüger et al. (2017). Second, we train on articles from multiple publishers and test on articles from an unseen publisher.

We demonstrate gains of using argumentation features on both collections and on all modeling approaches, with a wider margin of improvement in the smaller data scenario (i.e., when data from a single publisher is used in training).

[1]Covered in an ongoing shared task at SemEval 2023 task 3, subtask 1, where the task is to classify articles into news, opinions, or satire. More details in https://propaganda.math.unipd.it/semeval2023task3/index.html



(a) News Story

(b) Opinion Article

Figure 3.1: Sentences tagged as claims or premises in a news story and opinion articles.

3.1.1 Data

In our experiments on news vs. opinion classification, we use two data collections that aim to test the generalizability of the modeling approaches. Details about sizes, publishers, and dataset splits in both collections are shown in Table 3.1.

3.1.1.1 WSJ–NYT

For this dataset, we use the setup introduced by Krüger et al. (2017) for their work on news vs. opinion classification. This consists of data from two different publishers. From the BLIIP Wall Street Journal (WSJ) dataset (Charniak et al., 2000), we select 3,502 articles to create a balanced training set from the two classes, 1,751 news and 1,751 opinions (including editorials and letters to the editor), and a balanced test set of 1,000 articles from the WSJ. We create our datasets from the original WSJ corpus following the same approach described in Krüger et al. (2017), as the exact data splits are not publicly available. Finally, we use the New York Times Annotated (NYT) Corpus of the Linguistic Data Consortium (Sandhaus, 2008) to create two balanced sets of 2,000 articles each, one from the 'Armament, Defense and Military Forces' topic (henceforth NYT-Defense) and another one from the 'Medicine and Health' (henceforth NYT-Medicine) in order to measure the effect of publisher and topic shifts.

Data Collection	Type	Publisher	News	Opinion	Total
WSJ-NYT	train	WSJ	1751	1751	3502
	test	WSJ	500	500	1000
	test	NYT-Defense	1000	1000	2000
	test	NYT-Medicine	1000	1000	2000
Multi-Publisher	train	10 publishers	3193	3193	6386
	test	10 publishers	353	353	706
	test	The Metro - Winnipeg	418	418	836

Table 3.1: Details of all datasets from the two data collections.

3.1.1.2 Multi-Publisher

In order to understand the effect on this task when a model is trained on a diverse sample of articles, we create a data collection of 35k articles from multiple publishers. This collection consists of articles that are tagged as either regular news (90% of the data), or as opinions including op-eds, editorials, guests, letters, and others (10% of the data). The articles are from publishers in the US: *New York Times, Washington Post, Washington Observer Report, Digital Journal, Enid News, Californian, Press Democrat, NW Florida Daily, Gazette-Mail, and NJ Spotlight*. We split this data collection to train and test sets based on temporal information with the target of keeping a 90%-10% train-test split. We choose a date such that 90% of the articles in the data collection are published prior to that date and we consider those as the training split where the remaining 10% constitute the test split. Finally, we undersample the data by removing the extra news articles to have a balanced set of news and opinion articles.

The final training set consists of 6,386 articles and the final test set has 706 articles, all balanced across the two classes. We also create a balanced blind test set consisting of articles from an unseen publisher from Canada (The Metro-Winnipeg) totaling 836 articles crawled and undersampled in the same fashion. The majority of the articles in this data collection were published in 2018 or 2019.[2]

We perform preprocessing steps on all datasets by removing sentences with phrases such as

[2]This collection contains articles from publishers covered by LexisNexis at the time the research was done, or which have no collection restrictions for research purposes. Bloomberg provided the collection of URLs that make up the dataset.

"your article" or "your editorial" as they exclusively appear in opinion articles.

3.1.2 Features

We run our experiments on three feature sets testing all possible combinations among them.

3.1.2.1 Linguistic Features

We start with using linguistic features as presented in Krüger et al. (2017), as the claim is these generalize well across publishers and topics. We re-implement the set of linguistic features that performed the best in their experiments, namely: average sentence and token length (inverted), normalized frequencies of (negation, negation-suffix, digits, and interjection), ratios of ending character per sentence (question marks, exclamation points, commas, and semicolons), the ratio of quoted text, the ratio of verb tense outside the quoted text (past, present, future:will, modal verbs) of all verbs in the article, the sentiment of text outside quotes, the sentiment of adjectives outside quotes. We ignore features that require parsing to simplify feature extraction as they did not show significant gains in this task. The sentiment is represented by a numerical value that captures the degree ('weak-subj': 0.1, 'strongsubj': 1.0) and the polarity of the sentiment that is extracted using the MPQA Sentiment Clues Lexicon (Wilson et al., 2005). Our reproduction of Krüger et al. (2017) yields different results which are due to our more strict pre-processing that removes trivial cues from the data and a difference in how the dataset is sampled and split.

3.1.2.2 Document-level Contextualized Embeddings

We fine-tune *bert-base-cased* models (Devlin et al., 2019a) on each of the two data collections to obtain a contextualized representation of the article. We use the top layer of the [CLS] token to represent the article. We experiment with using each of the top four layers, the sum, and the average of all four layers to represent the [CLS]. The top layer yields the best results on the single publisher test sets with a small gain over other layers.

3.1.2.3 Argumentation Features

Since our target corpora do not have argumentative discourse unit (ADU) segmentation, we train a model to estimate argumentation features for each sentence in a news article. To this end, we use the corpus from Al Khatib et al. (2016) that has annotations of ADUs in 300 editorials from 3 publishers. Each ADU consists of one or more propositions and is annotated with one of six argumentative types:

- Assumption: states an assumption, conclusion, or opinion of the author that usually needs support by evidence.
- Common-Ground: states common knowledge or self-evident fact
- Testimony: gives evidence by quoting an authority
- Statistics: gives evidence by quoting results or conclusions of quantitative nature.
- Anecdote: gives evidence by stating an example.
- Other: Not argumentative or does not match any of the above types.

When training the model, we ignore sentences in the training data with multiple argumentative types among their propositions and assume one argumentative type span over a single sentence, similar to what is done in Daxenberger et al. (2017) where the claim detection task is structured as a sentence classification task. As our final objective is article-level classification, we expect this choice to have little effect on the downstream task.

We also group the six argumentative types into three coarser types, as some classes are infrequent or similar: claim (Assumption), premise (Common-Ground, Testimony, Statistics, Anecdote), and other (Other). We split the dataset into a training set of 6,316 sentences and a test set of 2,112 sentences. The training and the test sets are not balanced, where they have 65-70% claims, 30-35% premises, and only about 5% labeled as other. This is an important property of the writing style in editorials and will prove to be very useful for this task as we show in our results in Table

3.2. We train a BERT model (Devlin et al., 2019a) for 3 epochs (with the following values of the hyperparameters: 256 max sequence length, 32 training batch size, and 2e-5 learning rate) to perform a three-way sentence classification into claim, premise, or other. The classifier has a macro F1 score of 0.76 on the labeled test set. We experiment with other hyperparameters and other transformer-based models, such as RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), but notice negligible differences with respect to the fine-tuned BERT model.

We split the articles in all the datasets described in Section 3.1.1 into lists of sentences using the NLTK sentence tokenizer (Bird et al., 2009) and we use our fine-tuned BERT model to classify each sentence into one of the three argumentative types. We then use the tagged sentences in each article to generate argumentation features used in the main task of article-level news vs. opinion classification.

3.1.3 Models

We describe below the three models we use in our experiments, which include a machine learning model with discrete features, namely an SVM (Cortes and Vapnik, 1995), and two deep learning models, namely RNN (Rumelhart et al., 1985) and BERT.

SVM. We train a support vector machine (SVM) classifier with a linear kernel using scikit-learn implementation (Pedregosa et al., 2011). The SVM model can take as input the linguistic features, similar to the ones introduced by Krüger et al. (2017), the contextualized document representation generated by the BERT model, the argumentation features, or any combination of these. Argumentation features are represented as the distribution across the three classes (claims, premise, none) in a given article since our hypothesis is that editorials tend to have a majority of claim sentences, while news articles tend to have a majority of premises or other sentence types.

BERT. The BERT model is used to predict the type of article based on the [CLS] token that represents each article. BERT is implemented using the HuggingFace transformers library (Wolf et al., 2019). We train for 3 epochs, with a maximum sequence length of 512 tokens, a learning rate of

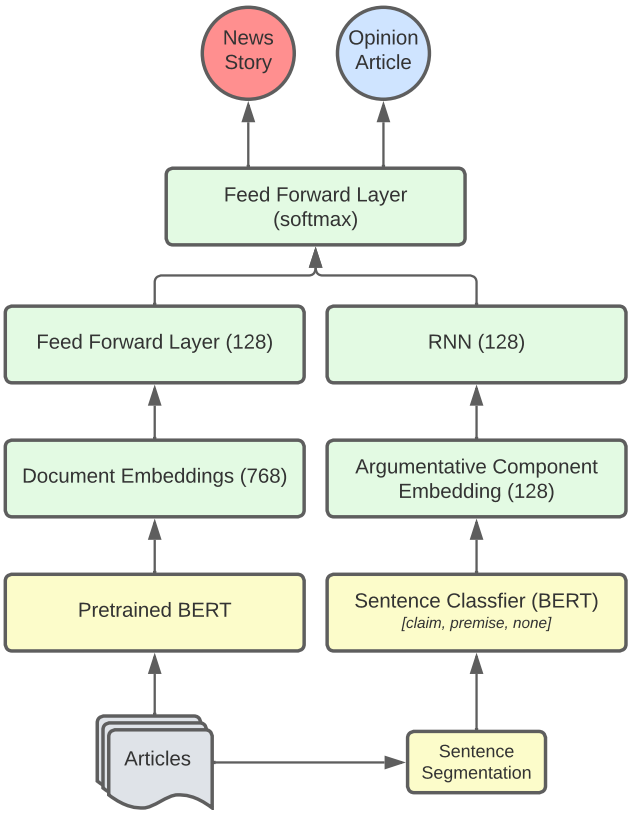


Figure 3.2: RNN+BERT model architecture.

2e-5, and a batch size of 16 on the training sets from both data collections.

RNN. We use a recurrent neural network (RNN) to bridge the gap between the sentence-level predictions for argumentation types and our document-level task of article classification. We hypothesize that the discourse relationships between the sentence-level predictions can be leveraged to improve classification when compared to only using the distribution over types.

For the RNN model, we use the argumentative labels of sentences as a sequence input to an RNN layer of size 128, with 20% dropout and 20% recurrent dropout for regularization. We pass the output of the RNN model to a softmax dense layer for prediction. The input sequence to the RNN has a maximum length of 100 sentences, which covers more than 95% of the articles. Since we have a sequence of a categorical feature that has one of three possible values (claim, premise, none) as opposed to a full vocabulary, we elected to use a vanilla RNN layer instead of more complex layers such as a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997)

or a Gated Recurrent Unit (GRU) (Cho et al., 2014).

RNN+BERT. In addition to fine-tuning BERT for the document-level task, we also use the fine-tuned BERT embeddings as input to a dense layer of size 128 with 50% dropout, and we concatenate the output with the RNN layer, then we pass the concatenated layer to a final softmax layer. We introduce a dense layer with a dropout after the BERT embeddings, such that the BERT and RNN output have equal layer sizes before concatenation. We denote this model as RNN+BERT. The diagram of the model is presented in Figure 3.2.

It is important to note that this work was done in 2020 while this thesis is written in 2022. More sophisticated language models were released in the past two years that can be used for this task such as BART (Lewis et al., 2020), DeBERTa (He et al., 2021), T5 (Raffel et al., 2020), T0 (Sanh et al., 2022), and GPT-3 (Brown et al., 2020). We leave experiments of these models on this task for future work.

3.1.4 Results

The results of our experiments on the WSJ–NYT collection are shown in Tables 3.2 and 3.4, while our results on the Multi-publisher collection are shown in Tables 3.3 and 3.5.

3.1.4.1 WSJ–NYT

All models are trained on the WSJ training set of articles that are classified as either news or opinion, where opinion articles include both editorials and letters to the editor. The results are shown in Table 3.2. The experiments uncover that using BERT pre-trained models in classification either by fine-tuning or by using their contextualized embeddings as features in an SVM model yields very high performance, but only for in-domain classification on the WSJ test set.

On the other hand, argumentation features perform the best on the two cross-publisher and cross-topic test sets (NYT-Defense, NYT-Medicine). Argumentation features consistently show good performance on all test sets both when used as aggregate features in the SVM model or as

sentence-level features in the RNN model. Using the argumentation features in the RNN model yields the highest performance on both of the NYT test sets, showing that modeling the discourse structure, rather than using aggregate distribution, is beneficial.

There is almost no effect from adding linguistic or argumentation features to embeddings. This could be due to the big difference in size between the 768-long feature vector of embeddings while other feature types have sizes less than twenty. To remedy the effect of feature sizes, we train an ensemble SVM model on the prediction probabilities from two SVM models: one with only BERT embeddings as features and another one with argumentation features only. This model performs better than embeddings-only; however, the ensemble model does not have the overall highest results on any of the test sets.

As mentioned in Section 3.1.2.1, we could not reproduce the results of using linguistic features exactly as described by Krüger et al. (2017) due to more strict pre-processing steps and different data splits. We notice this drop in performance when using argumentation features as well in our pilot experiments prior to using our more strict pre-processing steps that aim to remove trivial predictions. However, argumentation features show a smaller drop in performance caused by pre-processing (2-3 points in average F_1 score), which indicates their resilience to missing sentences from a given article.

The models using BERT representations have very high predictive performance when the test set is from the same publisher as the training set but generalize poorly to the other test sets from a different publisher (NYT) and on other topics (Defense and Medicine). We hypothesize this drop in performance may be caused by a lack of variety in the training data, which causes the model to learn representations that do not generalize well. The next set of experiments on the multi-publisher data collection studies the results of providing the model with data from a more varied set of publishers. Still, from training on a single publisher, we demonstrate that argumentation features transfer well to unseen publishers and topics without needing a large amount of task-specific training data.

Model	Features	WSJ	NYT-Def	NYT-Med
SVM	Ling.	0.84	0.75	0.70
	Emb.	**0.99**	0.79	0.78
	Arg.	0.89	0.88	0.87
	Ling. + Emb.	**0.99**	0.79	0.78
	Ling. + Arg.	0.91	0.88	0.87
	Emb. + Arg.	**0.99**	0.79	0.78
	ALL	**0.99**	0.79	0.78
SVM Ensemble	SVM Emb. SVM Arg	**0.99**	0.83	0.80
BERT	–	**0.99**	0.79	0.76
RNN	Arg.	0.94	**0.91**	**0.88**
RNN+BERT	Emb. + Arg.	**0.99**	0.79	0.78

Table 3.2: Average F_1 score for classification of articles into News or Opinion. All models are trained on a single publisher (WSJ). **NYT-Def**: Defense topic, **NYT-Med**: Medicine topic. **Bold**: highest overall. Underlined: highest in SVM only. All datasets are balanced.

Model	Features	Multi Publisher	Unseen Publisher
SVM	Emb	**0.93**	0.89
	Arg	0.84	0.89
	Emb+Arg	**0.93**	0.89
BERT	–	**0.93**	0.90
RNN	Arg	0.85	0.86
RNN+BERT	Arg+Emb	**0.93**	**0.91**

Table 3.3: Average F_1 score for classification of articles into News or Opinion. All models are trained on a the multi-publisher training data. All datasets are balanced.

3.1.4.2 Multi-Publisher

Table 3.3 presents the predictive results when training on the multi-publisher dataset and testing, separately, on data from the same publishers and the publisher unseen in training. Given that linguistic features did not do well on any of the test sets in the single-publisher training, we exclude them from our multi-publisher experiments.

The results show different patterns from the last experiment. In these settings, BERT or BERT-based features (in the SVM, or concatenated with the RNN) yield the best results on the multi-publisher test set. BERT is also able to generalize well on the unseen publisher in the test

set. However, the argumentation features used by the RNN model are still able to improve on the BERT results by 1 F1 point when used in combination with BERT. This shows that even with a more robust BERT classifier, the argumentation features can still improve the results on articles from the unseen publisher. Adding the argumentative feature also does not hurt performance when tested on the multi-publisher test set.

Remarkably, the argumentation features alone are able to achieve relatively high performance, despite the fact that they are of very low dimensionality and are trained on a distinct, albeit related task.

Examining the results from both the WSJ-NYT dataset and the multi-publisher training, we observe the ability of argumentation features to capture more global trends in the writing styles in news and opinion articles. Therefore, learning argumentation features from a single publisher proves to be enough to demonstrate good transferability across other publishers. This indicates that the global trends captured by the argumentation features are related to the structure of the article and its argumentative sentence types rather than specific phrases or topics used in the article.

On the other hand, BERT captures distinctive patterns related to the words, phrases, and topics used in the articles. This explains the large change in performance when trained on single or multiple publishers. This indicates the ability of BERT-based models to improve in terms of generalizability as the diversity of the training data increases. However, the argumentation features seem more suitable in data-scarce scenarios and can still add to rich BERT-based models trained on the task at hand.

3.1.4.3 Sub-types of Opinion Articles

To investigate the performance of argumentation features on specific types of opinion articles, we run experiments on two more tasks: news vs. editorial, and news vs. letters to the editor showing their results in Table 3.4 for the WSJ-NYT dataset and in Table 3.5 for the multi-publisher dataset. The results in Table 3.4 clearly show the advantage of using argumentation features in the editorial vs. news task. On the other hand, BERT performs better on the letters vs. news task which could be

Opinion Class	Dataset	SVM (Arg. features)	BERT	RNN
Editorial	NYT-Def	**0.90**	0.63	**0.90**
	NYT-Med	0.88	0.62	**0.91**
Letters	NYT-Def	0.89	**0.98**	0.87
	NYT-Med	0.88	0.85	**0.87**

Table 3.4: Average F_1 score for classification of news vs. editorials (top), and news vs. letters-to-the-editor (bottom). All models are trained on a single publisher (WSJ). All datasets are balanced.

Opinion Class	Dataset	SVM			BERT	RNN	RNN+BERT
		Emb	Arg	Emb+Arg	–	Arg	Arg+Emb
Editorial	Multi Publisher	0.93	0.90	0.93	**0.94**	0.89	0.91
	Unseen Publisher	0.89	0.88	0.88	0.89	0.87	**0.90**
Letters	Multi Publisher	0.98	0.86	0.98	**0.99**	0.89	0.95
	Unseen Publisher	**0.91**	0.88	**0.91**	**0.91**	0.87	0.87

Table 3.5: Average F_1 score for classification of news vs. editorials (top), and news vs. letters-to-the-editor (bottom). All models are trained on a the multi-publisher training data. All datasets are balanced.

due to the bigger lexical difference between these two types. Linguistic features from previous work also do well on classifying letters vs. news particularly due to the use of pronouns in the letters (Krüger et al., 2017). This is also true for the more resilient BERT model that is trained on multiple publishers (Table 3.5) where it performs better on the news vs. letters task. Similar to what we saw in the news vs. opinion task under multi-publisher training, the RNN+BERT model improves the results slightly over BERT on the news vs. editorial task when tested on the unseen publisher set.

3.1.5 Analysis of Argumentation Features

To further understand the relation between argumentative types of sentences and the discourse structure of the articles, we study the frequency of claims and premises at each sentence position. Figure 3.3 shows the number of times a claim (or a premise) is predicted at each sentence position normalized by the number of articles that have this sentence position, e.g., sentence 30 shows the number of times it is classified as a claim (or a premise) divided by the number of articles of length 30 or more. These percentages are calculated on the first 40 sentences from the articles in the

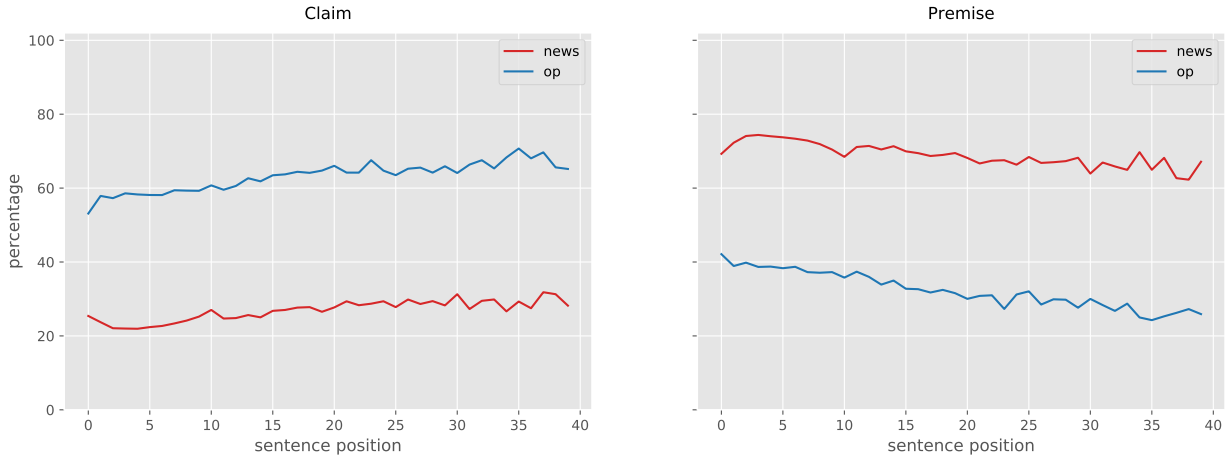


Figure 3.3: Frequencies of claims and premises at each sentence position in news and opinion articles.

multi-publisher training dataset in order to limit the variability caused by low counts.

Figure 3.3 shows that opinion articles tend to have a majority of claims and news articles tend to have a majority of premises, which explains the ability of simple features such as the distribution of sentence types to classify the article type as we show in Section 3.1.4.

In addition, we see a trend in the opinion articles to contain fewer premises, and conversely slightly more claims, as the article progresses. This trend is much less pronounced in news stories. These trends indicate that editorial and news stories follow, in aggregate, distinct discourse patterns. These differences in base rates justify why the SVM model using aggregate counts is able to predict with good accuracy the type of article with only a few features. In addition, by modeling more complex discourse dynamics across sentences by including the whole sequence of their argumentative types, the RNN model is able to further improve the performance when predicting document-level labels.

Editorials tend to have a majority of claims (assumptions) as mentioned by Al Khatib et al. (2016), which is consistent with our results. However, we see in our results that news articles tend to have a majority of premises, which could be the case for some but not all news articles. We think our model could be overestimating the number of premises in news articles due to being trained strictly on data annotated from editorials. In addition, some errors are caused by the sentence

segmentation model, which for example sometimes considers punctuations after abbreviations as sentence endings. Consider the news article in Figure 3.1a, we can see the sentence fragment "*A new law, signed Thursday by Gov.*" is marked as a complete sentence by the sentence segmentation model and classified as a premise. However, this example is neither a full sentence nor has any argumentative text. Other examples include tagging sentences as argumentative (claims or premises) even when the sentence segmentation is error-free. The total predictions of claims and premises constitute more than 95% of the predictions which is too high for texts that have both argumentative and non-argumentative discourse. The training data has a very small number of sentences from the 'non-argumentative' type and as a result, this class is under-predicted by the sentence-level model.

However, we believe that our predictions of sentence types are good estimates for article-level and possibly paragraph-level tasks, but more balanced training data from a diverse set of articles (editorials and news stories) is needed to apply this approach for sentence-level tasks.

3.2 Information Checkworthiness

We have seen how argumentation features help in distinguishing factual language from opinions in news articles. We now look at the problem of deciding what sentences to fact-check in news articles and in particular in the climate change domain. We hypothesize that selecting segments for fact-checking in news articles, particularly for controversial topics, is related to the overall argumentative structure of the article, more specifically to the argument component type (e.g., claim, premise) and to the incoming and outgoing argumentative relations (e.g., support, attack) from or to the argument components. By looking at some of the fact-checked articles, we notice that the segments selected for fact-checking by climate scientists sometimes contain a claim, a premise, or a combination of both a claim and a premise. When we look at the context around the fact-checked segments, we notice patterns related to the argumentative structure. For example, human fact-checkers tend to fact-check a claim when it is not supported by evidence (premise) or only supported by another claim, and fact-check a premise when it is used to support a claim (e.g.,

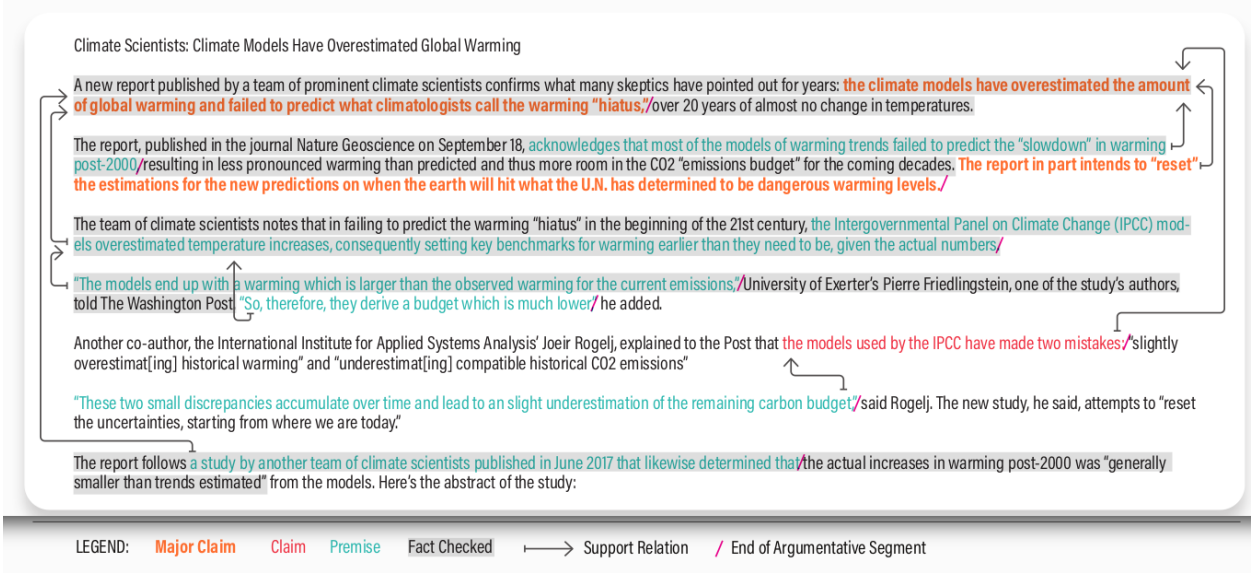


Figure 3.4: Fact-checked segments and argument components and relations in one article.

to challenge the relevance of that evidence in support of the claim). Not all fact-checked segments are chosen on a basis related to the argumentative structure, as we show in our analysis. However, annotations of fact-checked segments and argument component types allow us to understand and model this relation. Figure 3.4 shows an excerpt from one article in our dataset with its argument structure and fact-checked segment annotations.

3.2.1 Multi-Layer Annotated Corpus

We describe below the dataset with the fact-checked segment annotation by climate scientists and our annotation of the argumentative discourse structure on the same dataset.

3.2.1.1 Fact-Checked Segments Annotation

We introduce a new dataset of 95 climate change news articles fact-checked at the sentence level by climate scientists at the `climatefeedback.org` website. The articles are from 40 publishers mainly in the U.S., UK, and Australia (e.g., *The New York Times, The Guardian, The Washington Post, The Wall Street Journal, The Australian, The Telegraph, Forbes, USA Today, Breitbart,*

Credibility	very-low	very-low/low	low	neutral	high	high/very-high	very-high	mixed
Count	23	7	10	7	21	8	18	1

Table 3.6: Number of articles per credibility level.

and *Mashable*).[3] Each article is fact-checked by 3 to 5 climate scientists that evaluate scientific reasoning, add relevant information missed by the article and check for: factual accuracy, scientific understanding, logical reasoning, precision/clarity, sources quality, and fairness/objectivity[4]. The articles are given an article-level credibility assessment from very low to very high by the fact-checkers in addition to the segment-level annotation. Table 3.6 shows the number of articles in each of the eight degrees of credibility for news articles. The annotations of fact-checked segments vary in length from a fragment of a sentence to multiple sentences. We thus map these to binary labels at the sentence level: fact-checked sentences or non-fact-checked sentences. Each sentence is labeled as 'fact-checked' if it is fact-checked, has a fact-checked fragment, or is part of a multi-sentence fact-checked segment. We use the NLTK sentence segmenter (Loper and Bird, 2002) to split both the original articles and the fact-checked segments into a list of sentences.

A total of 134 articles were fact-checked by `climatefeedback.org` at the time of crawling this data (*May 2020*). However, we only include articles that have segment-level annotations and thus the final dataset has a total of 95 articles. We split the dataset into 68 articles in the training set (4,353 sentences in total, 824 are fact-checked), 7 articles in the development set (249 sentences in total, 55 are fact-checked), and 20 articles in the test set (970 sentences in total, 220 are fact-checked). We consider article credibility, publisher, and the ratio of fact-checked sentences when doing the split to make sure all data splits have articles from a diverse set of credibility levels, publishers, and styles. The ratio of fact-checked sentences in all three splits is 20-25% of the total number of sentences in the data.

[3]We collect the articles from LexisNexis, which licenses the use of data for research purposes.
[4]https://climatefeedback.org/process/

3.2.1.2 Argumentative Discourse Structure Annotation

We also annotate the argumentative discourse structure of the 95 fact-checked articles. Our annotation scheme is a slight modification of the one introduced by Stab and Gurevych (2017a). It has the following three types of argument components, each consisting of a single proposition:

- Major-Claim: a proposition that expresses the main stance the author takes about the main issue of the text.
- Claim: a stance relating to the main issue of the text that can support or undermine a major claim, or another claim.
- Premise: a proposition that expresses reasons to believe a given claim.

Also, our scheme identifies the four types of relations listed below. The relations are directed connections between components, such that each component may have no more than one outgoing relation.

- Support: occurs when a premise supports another premise, a claim, or a major-claim, or when a claim supports a major claim.
- Attack: occurs when a premise attacks another premise, a claim, or a major-claim, or when a claim attacks a major claim.
- Restate: indicates that two components of the same type (such as two claims) are the same (e.g., the author introduces a Major Claim and then restates it at the end of the article).
- Joint: occurs only between two adjacent premises and indicates that the two should be taken as a single argumentative unit. They are distinct propositions, but neither can be considered argumentative without the other.

Our annotation study consists of six annotators, all undergraduate students. We recruit annotators from the departments of Linguistics, English, and Comparative Literature. We train them on a

sample of articles, then assign each a 32-article batch. The articles are distributed such that each batch has three annotators. We use the Brat web server as our annotation tool.[5]

We create gold annotations for each article by synthesizing all three of its annotators' contributions. The text span for each gold component consists of the minimum common span of all overlapping components from the three annotations. We use majority voting to decide the label of the new gold component, with the label that occurs most often in the overlapping individual annotations being chosen as the gold label. In cases with a three-way tie between unlabelled, Premise, and Claim or Major-Claim, we determine the highest quality annotator of that span, where annotator quality is an ordinal ranking of all annotators in the study in descending order of their average pairwise agreement across all articles and use the label of the highest quality annotator. Once the gold argument components are created, we generate gold relations. First, we collect all outgoing relations from the individual annotators' components associated with a given gold argument component. We then remove any relations which begin or end at a component that was not included in the creation of a gold component. Then, for each gold argument component, we determine the gold relation by, in order of priority: adherence to guidelines, annotator quality, and the frequency with which the given relation type appears in our corpus. Adherence is a binary True or False depending on whether the proposed relation is consistent with our annotation schemes, such that an adherent relation is chosen when possible. To assess the quality of the resulting gold annotations, an expert meta-annotator then examined 18 of the resulting 95 annotated articles and recorded any instances in which they disagreed with the gold annotation. This comparison resulted in an agreement with the gold annotations 85.3% of the time.

We calculate inter-annotator agreement using two versions of dkpro-statistic's open-source [6] implementation of Krippendorff's alpha, which measures on a scale from -1 (inverse agreement) to 0 (agreement only by chance) to 1 (perfect agreement) (Bär et al., 2013; Krippendorff, 2011). When using the coding version, which uses only the labels assigned to each component, we find an overall inter-annotator agreement of .4368, with category agreements of .1745 for Premises,

[5] brat.nlplab.org
[6] dkpro.github.io/dkpro-statistics

.2175 for Claims, and .3782 for Major-Claims. Using the unitizing version, which takes into account both the label of each argument component and the span each annotator selected, we find an overall agreement of .2763, with agreements of .2803 for Premises, .2463 for Claims, and .4312 for Major-Claims. We also use the unitizing version to calculate each annotator's average pairwise overall agreement for the purpose of assessing annotator quality, finding a range from .1776 to .4641. The dataset comes from multiple publishers and countries and includes numerous types of articles such as editorials, op-eds, news analysis, and news reporting. This increases the complexity of the annotation task, which could explain the low Krippendorff's alpha scores for inter-annotator agreement.

3.2.2 Analysis of Argumentation in Fact-Checked Segments

To further understand the relation between argumentative discourse structure and fact-checked segments, we analyze the argument component types and relations of the fact-checked segments in the training data. To see the effect of our strategy in selecting gold argumentative spans and relations on the overlap with fact-checked segments, we do our analysis using the annotations of the best annotator for each article (overall highest in pairwise agreement with other annotators), and the gold annotations. We look at the original fact-checked segments before they are split to sentences as described in Section 3.2.1.1. This results in 589 fact-checked segments that mostly consist of multiple sentences (splitting them to sentences increases the number to 824 fact-checked sentences).

Argument Component Types. We first look at the best annotator's coding. Out of the 589 fact-checked segments, 430 maps to argument components in the articles. Out of argumentative fact-checked segments, 53% consist of a single argument component: 95 are Claims, 82 are Premises and 17 are Major-Claims, while the remaining consist of two (25%), three (10%), or four or more argument components (12%). Table 3.7 shows the most frequent argument component types of the fact-checked segments. When we use the gold annotations, the number of annotated segments in most articles decreases due to only including segments that are annotated by two or

Best Annotator		Gold Annotations	
AC Type	Frequency	AC Type	Frequency
Claim	110	Claim	91
Premise	100	Premise	76
Premise Premise	40	Major-Claim	22
Claim Claim	26	Premise Premise	20
Claim Premise	25	Claim Premise	17
Major-Claim	21	Claim Claim	12
Premise Claim	13	Premise Claim	9
Premise Premise Premise	10	Premise Claim Claim	4
Claim Claim Claim	8	Premise Premise Claim	4
Premise Claim Premise	7	Claim Premise Claim	4

Table 3.7: The most frequent argument component (AC) types of fact-checked segments.

Arg. Comp.	Best Annotator			Gold Annotations		
	Number and Type of Relations		Freq.	Number and Type of Relations		Freq.
Claim	1	$\xrightarrow{\text{sup}}$ Claim	18	1	$\xrightarrow{\text{sup}}$ Claim	12
	1	$\xrightarrow{\text{sup}}$ Major-Claim	13	1	$\xrightarrow{\text{sup}}$ Major-Claim	11
Premise	1	$\xrightarrow{\text{sup}}$ Claim	79	1	$\xrightarrow{\text{sup}}$ Claim	54
	2	$\xrightarrow{\text{att}}$ Claim $\xleftarrow{\text{sup/oth}}$ Premise	9	1	$\xrightarrow{\text{sup}}$ Premise	4
Major Claim	$\geq$5	$\xleftarrow{\text{sup}}$ Claim (all)	13	$\geq$4	$\xleftarrow{\text{sup}}$ Claim (all)	10
	1	$\xrightarrow{\text{oth}}$ Major-Claim	3	1	$\xrightarrow{\text{oth}}$ Major-Claim	2

Table 3.8: Relation counts for best annotator (left) and gold annotations (right).

more annotators. This reduces the argumentative fact-checked segments from 430 to 307 out of the 589 total fact-checked segments. This reduction cascades to the frequency of argument component types (Table 3.7) and relations counts (Table 3.8) in fact-checked segments.

Argumentative Relations. When we look at the relations from and to argument components that are fact-checked (as annotated by the best annotator), we notice that a Premise is fact-checked when it has one relation (mostly an outgoing support relation) and a Claim is fact-checked when it has many relations (up to four) with mixed directions (incoming, outgoing) and types (support, attack). This essentially maps to fact-checking a Premise when it is used as supportive evidence and fact-checking a Claim when it is central to the overall argument of the article. Also, Claims and

Major-Claims are fact-checked when they are only supported by other Claims (which could signal that the author is not providing evidence, thus showcasing an "*evading the burden of proof*" fallacy). An elaborate discussion of the fallacies in misinformation and beyond is provided in Chapter 4. The most frequent relation counts of fact-checked segments are shown in Table 3.8.

The general patterns found in the annotations of the best annotator still hold in the gold annotations. The only exception in the gold annotations is that a Major-Claim is fact-checked more often than segments consisting of two Premises or two Claims, which is mainly due to smaller overall counts of argument components (and relations) in the gold annotations.

3.2.3 Experimental Setup

We use the climate scientists' decision to fact-check a sentence as our gold labels for checkworthiness. In order to understand the capability of machine learning models to decide whether a sentence should be fact-checked, we introduce an experimental setup as follows. In line with previous work, we formulate this problem in two ways: a) **sentence classification task**, i.e. determining whether a given sentence should be fact-checked or not, and b) **sentence ranking** by checkworthiness. For the sentence classification task, we use Macro F1 scores as our evaluation metric, while for ranking we use Mean Average Precision (MAP). We experiment with fine-tuning BERT (Devlin et al., 2019a) using its implementation in the transformers library by HuggingFace (Wolf et al., 2020) with and without argumentation context as described below.

Baselines. We fine-tune BERT for 3 epochs (*bert-base-uncased*, max sequence length 256, batch size 16, learning rate 2e-5) using three different inputs to establish a baseline for this task. The first baseline is fine-tuning using only the target sentence for classification as the input (SENT). The other two configurations utilize the capability of BERT to handle two inputs. Therefore, we experiment with passing the target sentence with its previous sentence as input (PREV+SENT) and with its next sentence (SENT+NEXT). These two configurations essentially provide local **discourse context** following the natural order of sentences in the article.

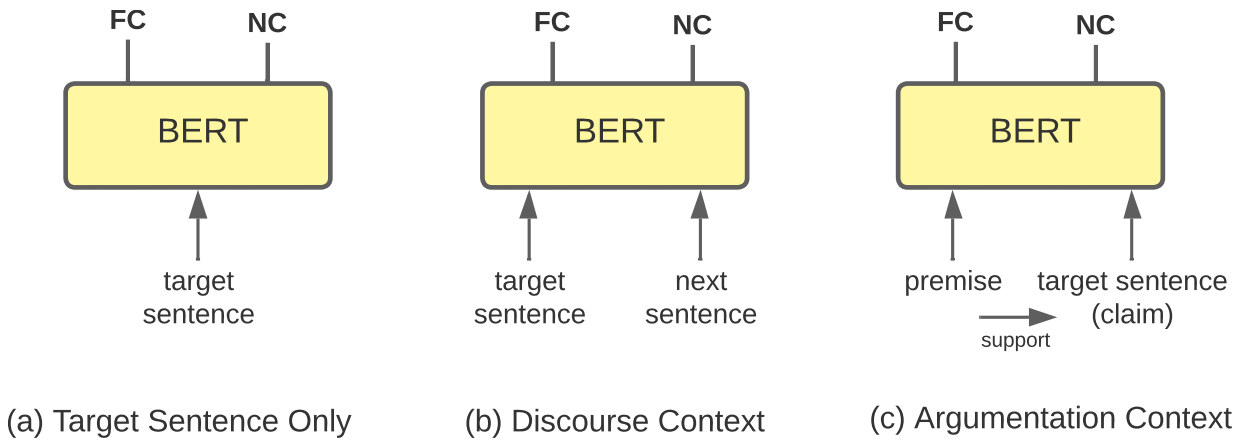


Figure 3.5: Showcasing three scenarios of fine-tuning BERT: (a) target sentence only, (b) an example of discourse context, and (c) and example of an argumentation context. The two labels are **FC**: Fact-Checked, and **NC**: Not Checked.

Argumentation Context. One simple way to test our hypothesis on the relation between argumentation and checkworthiness is by selecting a context for the target sentence using the argumentative discourse structure. We refer to such context as the argumentation context. If the target sentence is argumentative, we look at its outgoing and incoming argumentative relations. If the sentence has an incoming relation, then the source of that relation is passed as the first input of BERT and the target sentence is passed as the second input. If the relation is outgoing from the target sentence, then the target sentence is passed as the first input and the target of the relation is passed as the second. As a single sentence could consist of more than one argument component, which in turn could have many relations, this creates many pairs for the target sentence.

We explore three configurations for using the argument structure to select context. First, we keep all pairs for each target sentences, thus increasing the number of instances in the data and maintaining the same gold label for each repeated target sentence in the training data that is matched with a different argumentation context. We denote such configuration as AC(ALL). The final label during inference time can be determined in two ways: via majority label of predictions for each target sentence, and via favoring the minority class, i.e., if one prediction is to fact-check then we consider that as the final label.

Second, we select some of the argumentation context by keeping the most frequent relations

in fact-checked segments seen in training as discussed in Section 3.2.2. If the target sentence has a Claim or Major-Claim, then we only keep incoming support relations from other Claims or Major-Claims. However, if the target sentence has a Premise, we keep outgoing relations to Claims or Major-Claims. We also limit the total number by either 3 (AC(3)) or 1 (AC(1)) selecting at random if the remaining relations exceed the limit. In case the target sentence is not argumentative, we revert to the discourse context by selecting the previous sentence.

Third, we experiment with prepending the argument component type of the target sentence and its context to the input text (e.g., if the sentence has a claim, the input will be "_CLAIM_" followed by the sentence; for non-argumentative sentences we use "_NONE_"). We denote experiments with such configurations with the letter (T).

3.2.4 Results and Discussion

We show the results of our experiments in Table 3.9 for the development set and Table 3.10 for the test set. We can see in the baseline experiments in both tables that PREV+SENT condition is better than SENT+NEXT condition both in terms of Macro F1 score and the Fact-Checked class F1 score (FC_{class} F1).

Looking at the results on the dev set, we can see that the argument context of SENT+AC(1) has the highest FC_{class} F1 of 0.33, which is ***4 points*** above PREV+SENT and ***6 points*** above SENT+NEXT. It also has the highest Macro F1 of 0.58, which is ***2 points*** above PREV+SENT and ***3 points*** above SENT+NEXT. This indicates that providing a context based on argument relations that could be either before or after and not necessarily adjacent to the target segment is more informative for checkworthiness than providing local discourse context of the previous or next sentence. The same holds for the test set where the best argument context of SENT+AC(1)+T has the best FC_{class} F1 of 0.33 (***4 points*** above PREV+SENT and ***7 points*** above SENT+NEXT), best Macro F1 of 0.59 (***2 points*** above PREV+SENT and ***3 points*** above SENT+NEXT), and best MAP of 0.420 (***2 points*** above SENT, which is the highest baseline with MAP score). The test set SENT+AC(1)+T Macro F1 and MAP results are ***statistically significant*** over all three baselines SENT, PREV+SENT, and

Group	Model Input	Not-Checked	Fact-Checked	Macro F1	MAP
Baselines	SENT	0.83	0.23	0.53	0.296
	PREV+SENT	0.83	0.29	0.56	**0.387**
	SENT+NEXT	0.83	0.27	0.55	0.296
Argument Context (Text only)	SENT+AC(1)	0.84	**0.33**	**0.58**	0.366
	SENT+AC(3)^{v1}	0.82	0.31	0.57	0.299
	SENT+AC(3)^{v2}	0.82	0.32	0.57	0.299
	SENT+AC(ALL)^{v1}	0.83	0.26	0.54	0.318
	SENT+AC(ALL)^{v2}	0.81	0.30	0.56	0.318
Argument Context (Text+Type)	SENT+AC(1)+T	0.83	0.29	0.56	0.359
	SENT+AC(3)+T^{v1}	0.84	0.27	0.57	0.305
	SENT+AC(3)+T^{v2}	**0.85**	0.29	0.57	0.305
	SENT+AC(ALL)+T^{v1}	0.82	0.32	0.57	0.281
	SENT+AC(ALL)+T^{v2}	0.82	0.31	0.57	0.281

Table 3.9: Results on the development set. Per-class F1, macro F1 for sentence classification, and MAP for sentence ranking. ^{v1}Majority prediction to determine the final label. ^{v2}Final prediction is to fact-check if at least one prediction for the target sentence is as such. ^{v1,v2}Voting strategies do not affect MAP as we take the average of the prediction probabilities for each target sentence.

Input	NC	FC	F1	MAP
SENT	**0.85**	0.28	0.56	0.398
PREV+SENT	0.82	0.29	0.56	0.384
SENT+NEXT	0.84	0.26	0.55	0.385
SENT+AC(1)	0.83	0.30	0.57	0.413
SENT+AC(1)+T	0.84	**0.33**	**0.59**†	**0.420**†

Table 3.10: Per-class F1, macro F1 and MAP on the test set. †significant over the baseline (PREV+SENT).

SENT+NEXT.

However, providing more than one sentence does not improve the results in the AC(3) and AC(ALL) experiments as shown in Table 3.9, regardless whether the final prediction at inference time is decided via majority voting or favoring the FC class. Therefore, we only run AC(1) and AC(1)+T experiments on the test set. It is worth noting that adding the argumentative type to the target sentence and its context yields the highest results on the test set but not on the development set. This could be due to the small size of the development set of 249 sentences from 7 articles, which could have lead to high variability from the general trend in the data. The sentence type information has also the highest MAP score for the sentence ranking task. The ranking is done

based on the prediction probability of the model for all sentences in an article. The MAP value is computed by taking the mean of all average precision scores on all articles in one data split. This is a simplified version of the classification task where the model does not need to have correct prediction for every single sentence in the article as long as it highly ranks most of the fact-checked sentences in an article.

Argumentative Segments. In order to have a better understanding of the true potential of the argumentative discourse context for this task, we look at the accuracy of predictions on the argumentative segments of the articles. All non-argumentative segments have no incoming or outgoing argumentative relations. Therefore, there is no way of providing an argumentative discourse context for them, and thus they are matched with their previous sentence as mentioned earlier. Thus, the reported results on all AC conditions is on a mix of pairs where some sentences have an argumentation context while others have a discourse context. Out of the 249 sentences in the dev set, 133 are argumentative of which 37 are Fact-Checked. If we look at the model performance on this subset of the dev set, we see scores of 0.31 FC_{class} F1 and 0.53 Macro F1 for PREV+SENT, while having scores of 0.41 FC_{class} F1 and 0.60 macro F1 for SENT+AC(1). A gain of ***10 F1 points*** in the FC_{class} on the argumentative subset of the dev set compared with 4 points difference in FC_{class} F1 on the whole set shown in Table 3.9. The same observation holds for the test set that includes 485 argumentative sentences (out of 970) of which 123 sentences are Fact-Checked. The results on this subset are 0.33 FC_{class} and 0.55 macro F1 for PREV+SENT, and 0.38 FC_{class} and 0.61 macro F1 for SENT+AC(1)+T. This is again a wider margin of ***5 F1 points*** on FC_{class} compared to the 4 points difference in FC_{class} F1 reported in Table 3.10 on the whole test set. These numbers show that using argumentation context for determining the checkworthiness of sentences in an article is more clearly beneficial on the argumentative segments of the article. We leave further experimentation and modeling for future work that includes complimenting this approach with other linguistic information to determine the checkworthiness of the non-argumentative parts of the articles.

Error Analysis. We closely examine a few examples where the argumentation context has helped the model in making a correct prediction. One fact-checked "Major-Claim" says: *"Updated data from NASA satellite instruments reveal the Earth's polar ice caps have not receded at all since the satellite instruments began measuring the ice caps in 1979."* is the first sentence in the article, and thus it is paired with the title in the PREV+SENT model that does not make a correct prediction. However, the AC(1)+T model pairs it with another "Major-Claim" (*The updated data contradict one of the most frequently asserted global warming claims ...*) that comes 3 sentences later in the article and has a support relation to the target sentence. Another example is the "Major-Claim" (*The brutal weather has been supercharged by human-induced climate change*) that is supported by a "Claim" (*Climate models for three decades have predicted exactly what the world is seeing this summer*). Both of these examples are correctly predicted by the AC(1)+T model, which indicates the benefit of providing both the argument component type and its argumentation context to determine its checkworthiness, especially for "Major-Claims". On the other hand, AC(1)+T makes several wrong predictions to fact-check sentences from the Not-Checked class, which are predicted correctly by the SENT and the PREV+SENT models. This happens in cases where both the target and context sentences are Claim/Major-Claim, which indicates that such relations are providing a strong signal to fact-check. However, the climate scientists might have decided that those sentences are not checkworthy due to their own knowledge in the field rather than for reasons related to the argumentation structure.

3.3 Conclusion

Throughout this chapter, we have utilized knowledge from the argumentative discourse structure in two downstream tasks relevant to fact-checking. In the first task, we used a dataset of editorials with argument component annotations to train a sentence classifier that we used to extract features for the main target task of classifying news articles into news stories or opinion articles. In the second task, we annotated the argument structure in a dataset of fact-checked climate change news

articles and used these annotations to determine the most relevant context to provide for predicting the checkworthiness of sentences as determined by the human fact-checkers. Looking at our study in both tasks, we have the following observations:

- We are able to predict the argument structure to provide useful features for article-level classification, however, we had to use gold annotations of arguments to improve on sentence-level classification (and ranking).
- Argumentative discourse structures are able to provide highly transferable features under data scarce scenarios that are not affected by changes in domain and publisher.
- Preliminary results on studying the relation between argument structure and fact-checking show statistically significant improvements in utilizing argumentation context for predicting checkworthiness, especially in the argumentative segment of the target text.

Considering the aforementioned observations, we list two avenues of potential improvements.

- To have more accurate predictions of argument structures, we could train a multitask token and sentence argument segmentation model on news articles similar to our work on school student essays (Alhindi and Ghosh, 2021). End-to-End argumentation mining could also be improved further by incorporating information from boundaries of Elementary Discourse Units (EDUs) (Saha et al., 2022).
- It seems that argument structure improves checkworthiness prediction on a subset of check-worthy statements that follow our hypothesis: fact-checking a claim when it is not supported or only supported by other claims, and fact-checking a premise when it supports a claim. The first part of the hypothesis (fact-checking a claim) demonstrates the structural fallacy of *"Evading the Burden of Proof"*. In the next chapter, we study qualitative aspects of arguments and how they link to checkworthiness from a comprehensive view of fallacies that covers multiple types (beyond structural) and diverse datasets.

Chapter 4

Fallacies as Indicators of Misinformation

In the previous chapter, we showed how features from argument structure can help in downstream tasks such as the classification of article types, and information checkworthiness in news articles. In this chapter, we are interested in modeling qualitative aspects of arguments, more specifically fallacious moves in argumentation, and utilizing such models in deepening our understanding of misinformation and checkworthiness.

Fallacies are used as seemingly valid arguments to support a position and persuade the audience about its validity (Englebretsen, 1973). Theoretical work in argumentation has introduced various typologies of fallacies. For example, Van Eemeren et al. (2002) consider fallacies that occur when an argument violates the ten rules of a critical discussion (Van Eemeren and Grootendorst, 1987)[1], while Tindale (2007) thinks *"a precise definition (of fallacy) is difficult to give and depends on a range of considerations"* and categorizes fallacies into 4 categories: structural fallacies, related to the number and structure of arguments; fallacies from diversion, drawing from the (un)intentional diversion of the attention from the issue at hand; logical fallacies, related to the argument scheme at play and language fallacies, related to vagueness or ambiguity. Fallacious reasoning can bring misbehavior and be used for manipulation purposes. Thus, having a system that can find and classify fallacy types is crucial for applications that teach humans how to identify and avoid using fallacies in their arguments.

Our study of fallacy recognition in this chapter consists of four sections. In Section 4.1, we present an overview of three existing fallacy schemes (and datasets) that describe fallacy types in dialogue (Habernal et al., 2017), propaganda techniques in news (Da San Martino et al., 2019b),

[1]The ten rules and their violations are listed in Appendix A

and logical fallacies collected from educational websites (Jin et al., 2022). Then in collaboration with Musi et al. (2022), we introduce a new scheme for annotating fallacies in fact-checked content as well as release two new datasets built using this scheme.

Then, we introduce our models for fallacy tagging and classification. In Section 4.2, we work on a single fallacy dataset and introduce a BiLSTM-CRF sequence tagger that finds fallacious segments in news articles as well as classifies their types (Alhindi et al., 2019). Due to the complexity of the simultaneous tagging and classification of fallacy, we limit this study to one dataset, propaganda (Da San Martino et al., 2019b), that has eighteen propaganda techniques with annotation of fragment boundaries and type of propaganda techniques in news articles. We show the effect of one-hot encoded features from relevant dictionaries on having a high precision sequence tagger for a high multi-class classification task in a severely imbalanced dataset. We end Section 4.2 by showing a model for the classification of fallacy types (propaganda techniques) given propagandistic segments.

In an effort to make progress towards solving the general problem of fallacy recognition beyond a single dataset, in Section 4.3, we expand the classification aspect of the task by including five fallacy datasets from four fallacy schemes, but we limit the tagging aspect by providing fallacious segments only. For this, we propose a unified model based on multitask instruction-based prompting of the T5 model (Raffel et al., 2020) for generic fallacy type classification of 28 unique types given any sentence that contains a fallacious segment (Alhindi et al., 2022).

Finally, in Section 4.4, we use our model for fallacy type classification for understanding the checkworthiness of statements where we model fallacies as rationales that explain the reasons why certain statements must be prioritized for fact-checking.

Our main contributions in this chapter are as follows:

- We introduce a new scheme for annotating fallacy in misinformation and use that scheme to annotate fallacies in two domains: climate change, and Covid-19.

- We introduce instruction-based prompts to train a T5 multitask model for fallacy type classification across five datasets.

- We consider fallacy types as indicators of misinformation and use them to explain the checkworthiness of misinformative content.

4.1 Fallacy Datasets and Schemes

Work in computational models for fallacy detection is still in its infancy, with a limited set of relatively small datasets such as fallacies in question and answer dialog moves (Habernal et al., 2017); name-calling in social media debates (Habernal et al., 2018), fallacies as propaganda techniques in news (Da San Martino et al., 2019b); and logical fallacies from educational websites (Jin et al., 2022).

We describe in Section 4.1.1 three existing fallacy schemes (and datasets) for dialogue (ARGOTARIO) (Habernal et al., 2017), PROPAGANDA (Da San Martino et al., 2019b), and logical fallacies (LOGIC) (Jin et al., 2022) that we use in our experiments. Then, we introduce in Section 4.1.2 a new fallacy scheme for misinformation (MISINFO) in fact-checked content applied on text from two domains: climate change and Covid-19 (Musi et al., 2022; Alhindi et al., 2022).

Table 4.1, shows four examples of fallacies from these datasets. The four fallacy schemes identify different aspects of fallacies, and have different number of fallacy types (ARGOTARIO: 5, PROPAGANDA: 18, LOGIC: 13, MISINFO: 10).

4.1.1 Existing Fallacy Datasets and Schemes

4.1.1.1 ARGOTARIO

Introduced by Habernal et al. (2017), the Argotario dataset consists of five fallacies that appear in dialogue between players in game settings. The five fallacy types are: *Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization, irrelevant authority*, in addition to the *No Fallacy* type. The authors selected these types in particular because they are: common in argumentative discourse, distinguishable from each other, and have different difficulty levels. Players in the game are presented with a topic (question), which they answer using one of the fallacy types. Other

Question-Answering dialog moves in ARGOTARIO: Has anyone been on the moon? The moon is so far away, we should focus on our society. **Fallacy:** Red Herring
Propaganda techniques in news: The ability to build an untraceable, unregistered gun is definitely a game changer. **Fallacy:** Loaded Language
Educational website on fallacies: She is the best because she is better than anyone else **Fallacy:** Circular Reasoning
Fact-checked news: Says Joe Biden has said 150 million Americans died from guns and another 120 million from Covid-19. **Fallacy:** Cherry Picking

Table 4.1: Examples of fallacies from multiple datasets.

players then try to predict the fallacy type written by the author of the answer. The final label is determined when at least four players agree with the author of the answer on the type of fallacy. Each instance consist of a question-answer pair and one out of five fallacy labels.

4.1.1.2 PROPAGANDA

Propaganda aims at influencing a target audience with a specific group agenda using faulty reasoning and/or emotional appeals (Miller, 1939b). It is a form of communication that attempts to further the desired intent of the propagandist by emphasizing positive features and downplaying negative ones to cast an entity in a favorable light, which differs from persuasion that is interactive and attempts to satisfy the needs of both persuader and persuadee (Jowett and O'Donnell, 2012). Automatic detection of propaganda has been studied mainly at the article level (Rashkin et al., 2017a; Barrón-Cedeño et al., 2019). However, in order to build computational models that can explain why an article is propagandistic, the model would need to detect specific techniques present at the sentence or even the token level. Da San Martino et al. (2019b) identified the following 18 propaganda techniques that appear in news articles: *Loaded Language, Name Calling*

Propaganda Technique (*Fallacy Type*)	**Frequency**
Loaded Language	2,115
Name Calling,Labeling	1,085
Repetition	571
Doubt	490
Exaggeration,Minimisation	479
Flag-Waving	240
Appeal to Fear/Prejudice	239
Causal Oversimplification	201
Slogans	136
Appeal to Authority	116
Black-and-White Fallacy	109
Thought-terminating Cliches	79
Whataboutism	57
Reductio ad hitlerum	54
Red Herring	33
Bandwagon	13
Straw Men	13
Obfuscation,Intentional Vagueness,Confusion	11
Total	**6,041**

Table 4.2: Frequency of the eighteen propaganda techniques in the dataset.

or Labeling, Repetition, Exaggeration or Minimization, Doubt, Appeal to Fear/Prejudice, Flag-Waving, Causal Oversimplification, Slogans, Appeal to Authority, Black-and-White Fallacy, Thought-Terminating Cliche, Whataboutism, Reductio ad Hitlerum, Red Herring, Strawman, Bandwagon, and *Obfuscation, Intentional Vagueness, Confusion (OIVC).*

The data includes 350 articles in the training set, 61 articles in the development set, and 86 articles in the test set. The articles were taken from 48 news outlets; 13 propagandistic and 35 non-propagandistic as labeled by Media Bias/Fact Check[2]. These articles were annotated at the fragment level where each annotator was asked to tag the start and end of the propaganda text span as well as the type of propaganda technique. Table 4.2 lists all eighteen propaganda techniques and their frequencies in the training data. This is the biggest dataset in our experiments, but it is also the most unbalanced one, where six out of the 18 propaganda techniques represent more than 80% of all propagandistic segments. Each training instance consists of a sentence, a fragment, and one out of fourteen fallacy labels.

[2]https://mediabiasfactcheck.com/

4.1.1.3 LOGIC

Jin et al. (2022) collected examples of logical fallacies from educational websites on fallacies such as Quizziz, study.com and ProProfs. They identified 13 types of fallacies in the dataset using Wikipedia[3] as a reference. The fallacy types are: *Faulty Generalization, False Causality, Circular Claim, Ad Populum, Ad Hominem, Deductive Fallacy, Appeal to Emotion, False Dilemma, Equivocation, Fallacy of Extension, Fallacy of Relevance, Fallacy of Credibility* and *Intentional Fallacy*. Each training instance consists of a text segment (e.g., dialogue, sentence) and one of thirteen fallacy labels. The authors also introduce another challenge dataset: CLIMATELOGIC that follows the same fallacy scheme. However, it contains text segments that are too long (e.g., multiple paragraphs) with no annotations of smaller fallacious fragments like the Propaganda dataset. Therefore, CLIMATELOGIC is beyond the scope of this study.

4.1.2 A New Fallacy Scheme for Fact-Checked Content

While the three fallacy schemes in the previous section cover important aspects of fallacious moves in different scenarios, none of them are tailored towards fallacy types that are common in misinformation. In this section, we introduce a new annotation scheme for fallacy in fact-checked content that is developed in collaboration with Musi et al. (2022), as well as introduce two new fallacy datasets in the climate change and Covid-19 domains (Musi et al., 2022; Alhindi et al., 2022).

4.1.2.1 An Annotation Scheme of Fallacy

We adopt a bottom-up approach for developing the annotation scheme: an expert has analyzed 40 fact-checked articles randomly picked from the dataset and identified which fallacies have been called out through the comments of the fact-checkers. As an initial taxonomy of fallacies, we adopted the one proposed by Tindale (2007), which gathers the most common fallacies discussed in the informal logic tradition.

[3]https://en.wikipedia.org/wiki/List_of_fallacies

The resulting annotation schema includes ten types of fallacies related to:

- The argumentation structure:

  Evading the Burden of Proof

- The (un)intentional diversion of attention from the issue at hand:

  Strawman, False Authority, Red Herring, and *Cherry Picking*

- The argument schemes at play:

  False Analogy, Hasty generalization, Post Hoc, and *False Cause*

- The language used:

  Vagueness

As a reality check, we have analyzed the definitions of different "verdicts"/"labels" used by main fact-checkers in English (CLIMATEFEEDBACK; SNOPES; HEALTHFEEDBACK; POLITIFACT; FULLFACT; THEFERRET) to see whether critiques might point to fallacious moves different from the ones identified. Our set turned out to cover the fact-checkers verdicts: even if not exhaustive of the fallacy universe, our sub-selection is meant to represent the most frequent fallacious moves accomplished in online news. The guidelines contain a description of the notion of fallacy and its relation to fake news. Each fallacy is then defined, associated with an example, and accompanied by one or more critical questions, which have turned out to be useful means to evaluate arguments (Song et al., 2014). To offer systematic and economic heuristics, fallacies have been ordered starting from those having to do with the quantity of information provided (*structural fallacies*), followed by those related to aspects external to the issue discussed (*fallacies from diversion*); *logical fallacies* come into place after the other two classes are excluded. This order echoes the one provided by the pragma-dialectics rules for a critical discussion (Van Eemeren et al., 2002)[4], where the violations of rule 8 (Argument Scheme Rule) follow the violations of rule 2 (Burden-of-Proof Rule), rule 3 (Standpoint Rule) and rule 4 (Relevance Rule). It is, in fact, not worth looking at the argument

[4]Complete list of rules is provided in Appendix A

scheme at play if the information conveyed in the arguments is irrelevant to the conclusion. The *vagueness/ambiguity* fallacy occupies the last position in the heuristics when all the other options are excluded. We show below a list of all fallacy types along with their definitions.

- **Structural Fallacy** EVADING THE BURDEN OF PROOF: A position is advanced without any arguments supporting it as if it was self-evident
- **Fallacies from Diversion**
  - STRAWMAN: "The straw man fallacy is committed when the arguer misinterprets an opponent's argument for the purpose of more easily attacking it" (Hurley 1999, 119)
  - FALSE AUTHORITY: An appeal to authority is made where the source lacks credibility in the discussed matter, or they are attributed a statement which has been tweaked.
  - RED HERRING: The argument may be formally valid, but its conclusion is irrelevant to the issue at stake
  - CHERRY PICKING: The act of choosing among competing evidence that supports a given position, ignoring or dismissing findings that do not support it.
- **Logical Fallacies**
  - FALSE ANALOGY: "because two things [or situations] are alike in one or more respects, they are necessarily alike in some other respect" (Damer 1980: 49)
  - HASTY GENERALIZATION: A generalization is drawn from a numerically insufficient sample or a sample that is not representative of the population
  - POST HOC: It is assumed that because B happens after A, it happens because of A. In other words, a causal relation is attributed where, instead, a simple correlation is at stake.
  - FALSE CAUSE: X is identified as the cause of Y when another factor Z causes both X and Y OR X is considered the cause of Y when actually it is the opposite.

- **Language Fallacy** AMBIGUITY/VAGUENESS: A word, a concept, or a sentence structure which are ambiguous is shifted in meaning in the process of arguing or is left vague being potentially subject to skewed interpretations

We show an example below of a fact-checked segment along with the comment from the fact-checker followed by our ordered heuristics to annotate the fallacy.

- **Example:**
  - Segment: *Why is it that human emissions of carbon dioxide drive global warming yet natural emissions do not?*
  - Comment:“Nobody claims this"
- **Heuristics:**
  1. *Evading the Burden of Proof*:
     - CQ1: Does the position express an unassailable fact? No $\rightarrow$ CQ2
     - CQ2: Are there any arguments in support of the statement apart from personal guarantee? Yes$\rightarrow$ 2
  2. *Strawman*: Has an opponent’s position been misrepresented? Yes $\rightarrow$ END

4.1.2.2 Annotation

Using the same aricles mentioned in Section 3.2, we annotate 92 out of the 95 climate change articles that have a total of 735 fact-checked segments with comments from the fact-checkers. The annotators look at both the segment and comment when they annotate the fallacy type following the annotation scheme described in Section 4.1.2.1. The annotations were first done by two non-expert annotators that had a 0.47 Cohen’s κ (Cohen, 1960), which corresponds to moderate agreement. The gold labels were then done by an expert annotator (in argumentation theory) that went over both cases of agreement and disagreement to decide the final label. Following a similar setup, Musi

Fallacy Class	Fallacy Type	CLIMATE	COVID-19
None	No-Fallacy	389	613
Diversion	Cherry Picking (CP)	71	106
	Red Herring (RH)	46	35
	Strawman (S)	33	39
	False Authority (FAuth)	34	32
Logical	False Cause (FC)	36	17
	False Analogy (FA)	19	13
	Post-Hoc (PH)	13	20
	Hasty Generalization (HG)	7	78
Structural	Evading Proof (EBP)	28	105
Language	Vagueness (V)	59	77
All	All	735	1,135

Table 4.3: Fallacy statistics in the Climate Change and Covid-19 datasets.

et al. (2022) annotate 1,135 Covid-19 fact-checked claims that are found in social media, blogs, and news articles.[5]

We show the frequency of gold labels in both datasets of each fallacy type in Table 4.3. We notice in both datasets that around half of the fact-checked segments contain no fallacy. This indicates that these segments either present false information (*disinformation*) with no fallacious moves or are simply found to be true (*information*) after performing the manual fact-checking. When we look at the frequency of the fallacy types in each dataset, we notice some fallacies are among the most frequent in both domains such as *cherry-picking* and *vagueness*. We also see fallacies that are more frequent in Covid-19 (that come from a majority of social media posts) such as *hasty generalization*. The remaining of the diversion fallacies (*red herring, strawman,* and *false authority*) are moderately frequent in both datasets.

4.1.3 Unified Fallacy Types and Definitions

We list in Tables 4.4 and 4.5 all the definitions and fallacy labels used in all datasets. For the multitask (multi-dataset) model in Section 4.3, we unify the definitions and the labels for fallacies that fully or partially overlap. Additionally, in the same tables we show the original labels and definitions for all four fallacy schemes as they are released by (Habernal et al., 2017) for ARGOTARIO,

[5]Fact-checked by *Snopes, HealthFeedback, PolitiFact, FullFact, and TheFerret*

	Fallacy Type	Definition
(Habernal et al., 2017)	Ad Hominem	**The opponent attacks a person instead of arguing against the claims that the person has put forward.**
	Appeal to Emotion (Emotional Language)	**This fallacy tries to arouse non-rational sentiments within the intended audience in order to persuade.**
	Hasty Generalization	The argument uses a sample which is too small, or follows falsely from a sub-part to a composite or the other way round.
	Irrelevant Authority	While the use of authorities in argumentative discourse is not fallacious inherently, appealing to authority can be fallacious if the authority is irrelevant to the discussed subject.
	Red Herring	This argument distracts attention to irrelevant issues away from the thesis which is supposed to be discussed.
(Da San Martino et al., 2019b)	Black and White Fallacy	**Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an the extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).**
	Causal Oversimplification	**Assuming a single cause or reason when there are actually multiple causes for an issue.**
	Doubt	**Questioning the credibility of someone or something.**
	Exaggeration or Minimization	**Either representing something in an excessive manner: making things larger, better, worse or making something seem less important than it really is**
	Appeal to fear/prejudice (Fear or Prejudice)	**Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases the support is based on preconceived judgements.**
	Flag-Waving	**Playing on strong national feeling (or to any group) to justify/promote an action/idea.**
	Appeal to Authority (Irrelevant Authority)	Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert in this technique, although it is referred to as Testimonial in literature.
	Loaded Language	**Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.**
	Name Calling or Labeling	**Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises.**
	Red Herring	Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.
	Reductio Ad Hitlerum	**Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.**
	Slogans	**A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.**
	Thought-Terminating Cliches	**Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or distract attention away from other lines of thought.**
	Whataboutism	**A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.**

Table 4.4: Fallacy types and definitions (part 1).

(Jin et al., 2022)	Ad Hominem	An irrelevant attack towards the person or some aspect of the person who is making the argument, instead of addressing the argument or position directly.
	Ad Populum	**A fallacious argument which is based on affirming that something is real or better because the majority thinks so.**
	False Dilemma (Black and White Fallacy)	A claim presenting only two options or sides when there are many options or sides.
	False Causality (Causal Oversimplification)	A statement that jumps to a conclusion implying a causal relationship without supporting evidence
	Circular Reasoning	**A fallacy where the end of an argument comes back to the beginning without having proven itself.**
	Deductive Fallacy	**An error in the logical structure of an argument.**
	Appeal to Emotion (Emotional Language)	Manipulation of the recipient's emotions in order to win an argument.
	Equivocation	**An argument which uses a phrase in an ambiguous way, with one meaning in one portion of the argument and then another meaning in another portion.**
	Fallacy of Extension	**An argument that attacks an exaggerated/caricatured version of an opponent's.**
	Faulty Generalization (Hasty Generalization)	An informal fallacy wherein a conclusion is drawn about all or many instances of a phenomenon on the basis of one or a few instances of that phenomenon is an example of jumping to conclusions.
	Intentional Fallacy	**Some intentional/subconscious action/choice to incorrectly support an argument.**
	Fallacy of Credibility (Irrelevant Authority)	An appeal is made to some form of ethics, authority, or credibility.
	Fallacy of Relevance (Red Herring)	Also known as red herring, this fallacy occurs when the speaker attempts to divert attention from the primary argument by offering a point that does not suffice as counter-point/supporting evidence (even if it is true).
(Musi et al., 2022)	Evading the Burden of Proof	**A position is advanced without any arguments supporting it as if it was self-evident.**
	Cherry Picking	**The act of choosing among competing evidence that which supports a given position, ignoring or dismissing findings which do not support it.**
	Red Herring	**The argument supporting the claim diverges the attention to issues which are irrelevant for the claim at hand.**
	Strawman	**When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.**
	False Authority (Irrelevant Authority)	**An appeal to authority is made where the it lacks credibility or knowledge in the discussed matter or the authority is attributed a tweaked statement.**
	Hasty Generalization	**A generalization is drawn from a sample which is too small, not representative of the population or not applicable to the situation if all the variables are taken into account.**
	False Cause (Causal Oversimplification)	X is identified as the cause of Y when another factor Z causes both X and Y OR X is considered the cause of Y when actually it is the opposite
	Post Hoc (Causal Oversimplification)	It is assumed that because B happens after A, it happens because of A. In other words a causal relation is attributed where, instead, a simple correlation is at stake
	False Analogy	**because two things [or situations] are alike in one or more respects, they are necessarily alike in some other respect.**
	Vagueness	**A word/a concept or a sentence structure which are ambiguous are shifted in meaning in the process of arguing or are left vague being potentially subject to skewed interpretations.**

Table 4.5: Fallacy types and definitions (part 2).

	Fallacy	ARGOTARIO train/dev/test	PROPAGANDA train/dev/test	LOGIC train/dev/test	COVID-19 train/dev/test	CLIMATE train/dev/test	Total train/dev/test	Total All
1	Ad Hominem	102 /26/ 31	–/–/–	406 /64/ 81	–/–/–	–/–/–	508 /90/ 112	710
2	Ad Populum	–/–/–	–/–/–	296 /81/ 62	–/–/–	–/–/–	296 /81/ 62	439
3	B&W Fallacy	–/–/–	60 /16/ 19	192 /40/ 25	–/–/–	–/–/–	252 /56/ 44	352
4	Causal Ov.simp.	–/–/–	111 /28/ 34	303 /49/ 36	36 /10/ 10	39 /10/ 11	489 /97/ 91	677
5	Cherry Picking	–/–/–	–/–/–	–/–/–	76 /20/ 23	67 /17/ 21	143 /37/ 44	224
6	Circular Reason.	–/–/–	–/–/–	238 /40/ 35	–/–/–	–/–/–	238 /40/ 35	313
7	Deductive	–/–/–	–/–/–	205 /28/ 31	–/–/–	–/–/–	205 /28/ 31	264
8	Doubt	–/–/–	263 /66/ 82	–/–/–	–/–/–	–/–/–	263 /66/ 82	411
9	Emotional Lang.	150 /38/ 47	–/–/–	230 /38/ 41	–/–/–	–/–/–	380 /76/ 88	544
10	Equivocation	–/–/–	–/–/–	62 /13/ 11	–/–/–	–/–/–	62 /13/ 11	86
11	Evad Burd Prf	–/–/–	–/–/–	–/–/–	76 /20/ 23	31 /8/ 9	107 /28/ 32	167
12	Exag/Mini	–/–/–	304 /76/ 94	–/–/–	–/–/–	–/–/–	304 /76/ 94	474
13	Extension	–/–/–	–/–/–	187 /31/ 46	–/–/–	–/–/–	187 /31/ 46	264
14	False Analogy	–/–/–	–/–/–	–/–/–	13 /5/ 3	17 /5/ 5	30 /10/ 8	48
15	Fear/Prejudice	–/–/–	131 /33/ 41	–/–/–	–/–/–	–/–/–	131 /33/ 41	205
16	Flag-Waving	–/–/–	145 /37/ 45	–/–/–	–/–/–	–/–/–	145 /37/ 45	227
17	Hasty General.	104 /26/ 32	–/–/–	561 /128/ 123	54 /15/ 16	4 /2/ 2	723 /171/ 173	1,067
18	Intentional Fal.	–/–/–	–/–/–	215 /34/ 26	–/–/–	–/–/–	215 /34/ 26	275
19	Irrelevant Auth.	92 /24/ 29	57 /15/ 17	196 /18/ 33	26 /8/ 8	32 /8/ 10	403 /73/ 97	573
20	Loaded Lang.	–/–/–	1,331/333/416	–/–/–	–/–/–	–/–/–	1,331/333/416	2,080
21	Name Calling	–/–/–	685 /172/ 214	–/–/–	–/–/–	–/–/–	685 /172/ 214	1,071
22	Red Herring	115 /29/ 35	16 /4/ 10	214 /43/ 46	28 /8/ 8	44 /12/ 13	417 /96/ 112	625
23	Reductio AH.	–/–/–	33 /9/ 10	–/–/–	–/–/–	–/–/–	33 /9/ 10	52
24	Slogans	–/–/–	84 /22/ 26	–/–/–	–/–/–	–/–/–	84 /22/ 26	132
25	Strawman	–/–/–	4/1/6	–/–/–	28 /8/ 8	23 /6/ 7	55 /15/ 21	91
26	Thought-Term.	–/–/–	48 /12/ 14	–/–/–	–/–/–	–/–/–	48 /12/ 14	74
27	Vagueness	–/–/–	–/–/–	–/–/–	53 /15/ 23	48 /12/ 14	101 /27/ 37	165
28	Whataboutism	–/–/–	33 /9/ 10	–/–/–	–/–/–	–/–/–	33 /9/ 10	52
	Total (tr/de/te)	563 /143/ 174	3,305 /833/ 1,038	3,305 /607/ 596	390 /109/ 122	305 /80/ 92	7,868 /1,772/ 2,022	
	Total (All)	880	5,176	4,508	621	477	11,662	

Table 4.6: Counts of fallacy types in each split across all datasets.

(Da San Martino et al., 2019b) for PROPAGANDA, (Jin et al., 2022) for LOGIC, and (Musi et al., 2022) for MISINFORMATION that is used for the COVID-19 and CLIMATE datasets.

For the multitask prompting-based generative model (Section 4.3), we unify the labels of similar fallacies (e.g., *False Cause*, *False Causality*, *Causal Oversimplification* → *Causal Oversimplification*; *False Authority*, *Appeal to Authority*, *Fallacy of Credibility*, *Irrelevant Authority* → *Irrelevant Authority*). We also rephrase some fallacy types by removing words such as "Appeal to" (e.g., *Appeal to Emotion* → *Emotional Language*) that tend to throw off generative models causing over-prediction of these types as observed in our initial experiments. Some fallacies have partial or full overlap with others across the four schemes. Therefore, we merge these types and use the label of the most frequent or the most representative label of the fallacy type (e.g., *Fallacy of Relevance* → *Red Herring*; *Post Hoc* → *Causal Oversimplification*; *False Dilemma* → *Black-and-White Fallacy*). We also unify the definitions of fallacy types in prompts across datasets. We have a total of 28

unique fallacy types across five datasets with the following final number of fallacy types per scheme after merging and filtering out some of the types:

- ARGOTARIO: 5 → 5
- LOGIC: 13 → 13
- PROPAGANDA: 18 → 14
- MISINFORMATION: 10 → 9

The above fallacy types add up to 41 with 13 of them repeated across schemes (e.g., *Irrelevant Authority* and *Red Herring* exist in all four schemes). Therefore, we end up with 28 unique fallacies. We show counts of fallacy types in training/dev/test splits for all datasets in Table 4.6.

4.2 Single Dataset Case: Sequence Tagging and Fallacy Type Classification

We first consider a single fallacy dataset and start with the problem of simultaneous tagging and classification of fallacy types in Section 4.2.1 followed by our work on type classification of fallacious segments in Section 4.2.2.

In Section 4.2.1, a model is given a news article and needs to detect all spans of the text in which a fallacy occur. In addition, for each span the fallacy type must be identified. Due to the complexity of this task and the need for token level annotations, we only experiment with one fallacy dataset: PROPAGANDA. We participated in the 2019 shared task organized by Da San Martino et al. (2019a) for detecting 18 propaganda techniques at the fragment level (Alhindi et al., 2019). We study the ability of models to both find a fallacious segment as well as classify its type. This setup mimics real-life scenarios in which a system is not given a fallacious segment but rather tasked with finding one from a collection of text and classifying the type of fallacy it has.

The complexity of the task is exacerbated by the severe imbalance nature of fallacy datasets where a single "No Fallacy" class is more frequent than all 10-18 fallacy types combined (e.g. 70% of PROPAGANDA and 50% of COVID-19 are labeled as "No Fallacy"). Therefore, we remove the effect of the "No Fallacy" by only considering propagandistic segments in Section 4.2.2 to study the ability of models to classify the type of fallacy in a given piece of text. This setup is in line with

the propaganda technique classification task (Da San Martino et al., 2020) and the logical fallacy detection task (Jin et al., 2022) that do not include "No Fallacy" class.

4.2.1 Sequence Tagging

We use all 18 propaganda techniques in a joint tagging and classification task of propaganda techniques in Section 4.2. The model in this task is given a sentence and is asked to tag each word with (B/I)-propaganda technique or O. We divided the training set (350 articles) into a training set of 280 articles and a development set of 70 articles to perform ablation studies and error analysis.

4.2.1.1 Method

Our architecture builds on the FLAIR framework (Akbik et al., 2018; Akbik et al., 2019) that combines character level embeddings with different kinds of word embeddings as input to a BiLSTM-CRF model (Ma and Hovy, 2016; Lample et al., 2016). Akbik et al. (2018) have shown that stacking multiple pre-trained embeddings as input to the LSTM improves the performance on the downstream sequence labeling task. We combine GloVe embeddings (Pennington et al., 2014) with Urban Dictionary[6] embeddings[7].

We additionally include one-hot-encoded features based on dictionary look-ups from the UBY dictionary provided by Gurevych et al. (2012). These features are based on concepts associated with the specific word such as *offensive, vulgar, coarse,* or *ethnic slur*. In total, 30 concept features were added as additional dimensions to the embedding representations.

We also experimented with stacking BERT embeddings with all or some of the embeddings mentioned above, but it did not yield better results than the BiLSTM-based model. The best model used urban-GloVe embeddings with concatenated one-hot encoded UBY features stacked with both forward and backward FLAIR embeddings. The model was trained for a maximum of 150 epochs with early stopping using a learning rate of 0.1, a batch size of 32, and a BiLSTM with hidden size

[6]https://www.urbandictionary.com/

[7]https://data.world/jaredfern/urban-dictionary-embedding

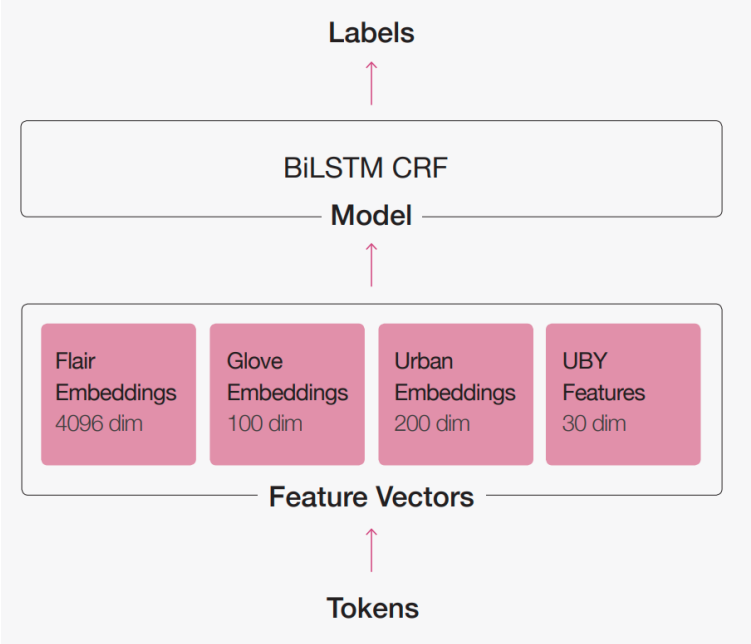


Figure 4.1: BiLSTM-CRF model with embeddings and handcrafted features

256. The results of this model are shown in Table 4.7.

4.2.1.2 Task Setup and Evaluation

This task follows a traditional `BIO` tagging scheme for name entity recognition (NER) that consists of Beginning of entity (B), Inside-entity (I), and Outside-entity (O) (Ramshaw and Marcus, 1999). However, propaganda techniques differ from typical NER tasks in the following: (i) techniques can overlap, and (ii) some techniques have long spans that cover a complete sentence or longer. Da San Martino et al. (2019b) develop a suitable evaluation metric for this task that is derived from the NER literature (Nadeau and Sekine, 2007), and other tagging tasks that are more similar to propaganda techniques such as plagiarism detection (PD) (Potthast et al., 2010).

This task is evaluated based on the prediction of the type of propaganda technique and the intersection between the gold and the predicted spans. For a document d represented as a sequence of characters with a set of gold propaganda fragments G that can possibly overlap, then a gold fragment $t = [t_i, ..., t_j] \subseteq d$ and a predicted fragment $s = [s_m, ..., s_n] \subseteq d$ are both associated with one of the eighteen techniques through a labeling function $l(x) = \{1, ..., 18\}$. Then, the precision

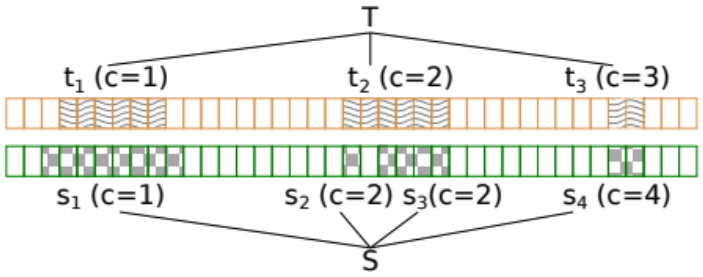


Figure 4.2: Evaluation measure for propaganda tagging (Da San Martino et al., 2019b).

and the recall are calculated as follows

$$P(S,T) = \frac{1}{|S|} \sum_{s \in S,\ t \in T} C(s,t,|s|)$$

$$R(S,T) = \frac{1}{|T|} \sum_{s \in S,\ t \in T} C(s,t,|t|)$$

where $C(s,t,h) = \frac{|s \cap t|}{h} \delta(l(s), l(t))$, and $\delta(a,b) = 1$ if $a = b$, and 0 otherwise.

4.2.1.3 Results and Analysis

We only show the results of our best model in Table 4.7 to focus more on the differences between propaganda techniques. The best model is a BiLSTM-CRF with FLAIR and urban GloVe embeddings with one hot encoded features as mentioned in Section 4.2.1.1.

As we can see in Table 4.7, we can divide the propaganda techniques into three groups according to the model's performance on the development and test sets. The first group includes techniques with non-zero F1 scores on both datasets: *Flag-Waving*, *Loaded Language*, *Name Calling,Labeling*, and *Slogans*. This group has techniques that appear frequently in the data and/or techniques with strong lexical signals (e.g., "American People" in *Flag-Waving*) or punctuation signals (e.g., quotes in *Slogans*). The second group has the techniques with a non-zero F1 score on only one of the datasets but not the other, such as *Appeal to Authority*, *Appeal to Fear*, *Doubt*, *Reduction*, and *Exaggeration,Minimisation*. Two out of these five techniques (*Appeal to Fear* and *Doubt*) have very small non-zero F1 on the development set, which indicates that they are generally challenging on

Propaganda Technique (*Fallacy Type*)	**Development**			**Test**
	P	**R**	**F**	**F**
Appeal to Authority	0	0	0	0.212
Appeal to Fear/Prejudice	0.285	0.006	0.011	0
Bandwagon	0	0	0	0
Black-and-White Fallacy	0	0	0	0
Causal Oversimplification	0	0	0	0
Doubt	0.007	0.001	0.002	0
Exaggeration,Minimisation	0.833	0.085	0.154	0
Flag-Waving	0.534	0.102	0.171	0.195
Loaded Language	0.471	0.160	0.237	0.130
Name Calling,Labeling	0.270	0.112	0.158	0.150
Obfuscation, Intentional Vagueness, Confusion	0	0	0	0
Red Herring	0	0	0	0
Reductio ad hitlerum	0.318	0.069	0.113	0
Repetition	0	0	0	0
Slogans	0.221	0.034	0.059	0.003
Strawman	0	0	0	0
Thought-terminating Cliches	0	0	0	0
Whataboutism	0	0	0	0
Overall	0.365	0.073	0.122	0.131

Table 4.7: Precision, recall and F1 scores of the FLC task on the development and the test sets

our model and were only tagged due to minor differences between the two datasets. However, the remaining three types show significant drops from the development to the test sets or vice-versa. This requires further analysis to understand why the model was able to do well on one dataset but got zero on the other dataset, which we leave for future work. For the remaining nine techniques, our sequence tagger fails to correctly tag any text span in either dataset. These techniques have the most infrequent types as well as types that are beyond the ability of our tagger to spot by looking at the sentence only such as *Repetition*.

In general, the model manages to detect, with varying levels of accuracy, fallacies that have emotional manipulations, short spans of texts, and strong lexical cues. However, it misses other fallacies that tend to be longer in nature and use a logical connection (e.g, *Causal Oversimplification*) or diversion (e.g., *Strawman*). Overall, our model has the highest precision among all teams on both datasets, which could be due to adding the UBY one-hot encoded features that highlighted some strong signals for some propaganda types. This also could be the reason for our model to have the

lowest recall among the top 7 teams on both datasets as having explicit handcrafted signals suffers from the usual sparseness that accompanies these kinds of representations which could have made the model more conservative in tagging text spans.

We noticed two types of noise in the data; there were some duplicate articles, and in some articles, the ads were crawled as part of the article and tagged as non-propaganda. These could have caused some errors in predictions and therefore investigating ways to further clean the data might be helpful.

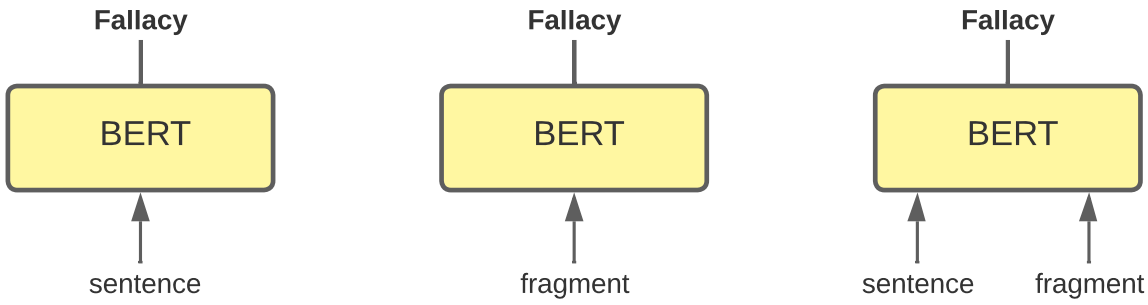
4.2.2 Fallacy Type Classification

Given the complexity of the task, we focus now on the classification aspect of fallacy recognition where a model is given a fallacious segment and asked to determine the type of fallacy it has. The data has annotations of the text spans of propaganda techniques (fallacy type) in 451 articles from 48 news outlets allowing multiple labels and partial overlap of text spans. We frame this as a sentence classification task and a fragment (part of a sentence) classification task.

4.2.2.1 Model and Experimental Setup

We only include propagandistic fragments or their containing sentences following a similar setup of the propaganda technique classification task introduced by Da San Martino et al. (2020). However, we remove the *Repetition* class since it requires a larger context (e.g., the full article) and does not have an argumentative fallacy. We also ignore propaganda fragments that span across multiple sentences. Considering the sentence level, the fallacy type becomes the label of the sentence if the fragment is included within the sentence. For sentences with multiple fragments, we consider the label of the longer fragment and we do not allow multiple labels for a single sentence. The data has more than 5k sentences with fallacious (propagandistic) fragments. The training, development, and test splits are shown in Table 4.6.

We fine-tune BERT for the fallacy type (propaganda technique) classification task under three conditions. First, sentence classification by providing sentences that contain fallacious fragments.



(a) Sentence containing a fallacious fragment

(b) Fallacious fragment

(c) Fallcious fragment and its containing sentence

Figure 4.3: Fine-tuning BERT for fallacy type (propaganda technique) classification.

Second, fragment classification by only providing the fallacious fragment that varies in length from a single word to a full sentence. Finally, we utilize the capability of BERT to handle two inputs and fine-tune the model by providing both the fallacious fragment and its containing sentence. We show the three conditions for fine-tuning BERT in Figure 4.3.

4.2.2.2 Results and Analysis

We show the results of fine-tuning BERT for the classification of fallacy types in Table 4.8 using the following hyperparameters: 3 epochs, 2e-5 learning rate, batch size 16, and maximum sequence length 256.

We can notice the big difference in scores between sentence-only and fragment-only classification across all types. This is most likely due to the short fragment lengths for some fallacy types (e.g., "right-wing Islamophobes" for *Name Calling* and "America First" for *Flag-Waving*), and the long fragments for others (e.g., "Did he know something that X was going to do?" for *Doubt*), which provides an important feature for classification. However, providing full sentences with no fragments increases the complexity of the classification task. The results improve when both the fragment and the sentence are used as inputs (Sent+Frag) in comparison with sentence-only classification (Sent), but not as high as fragment-only classification (Frag) as shown in Table 4.8.

The results indicate the capability of BERT to classify fallacy types when fallacious fragments

Propaganda Technique (*Fallacy Type*)	Sent	Frag	Sent +Frag
Appeal to Authority	0.13	0.29	0.30
Appeal to Fear/Prejudice	0.32	0.45	0.29
Bandwagon	0	0	0
Black-and-White Fallacy	0.07	0.30	0.09
Causal Oversimplification	0.20	0.45	0.42
Doubt	0.43	0.66	0.62
Exaggeration,Minimisation	0.34	0.56	0.54
Flag-Waving	0.49	0.58	0.61
Loaded Language	0.65	0.82	0.82
Name Calling,Labeling	0.51	0.81	0.82
Obfuscation, Intentional Vagueness, Confusion	0	0	0
Red Herring	0	0	0
Reductio ad hitlerum	0.18	0.46	0.37
Slogans	0.13	0.64	0.57
Strawman	0	0	0
Thought-terminating Cliches	0.09	0	0.30
Whataboutism	0	0	0.18
Accuracy	0.50	0.70	0.69
Macro F1	0.21	0.40	0.35

Table 4.8: Fallacy type classification F1 scores using BERT. **Sent:** sentence containing a propagandistic fragment. **Frag:** propagandistic fragment.

are annotated by having a 70% accuracy and a 40% macro F1 score. However, how can we improve the results when such fine-grained annotations are not available or when the number of fallacy types increases to cover ones that exist in different domains and genres? We address these questions in the next section by developing a unified model for fallacy recognition through multitask instruction-based prompting.

4.3 Multi Dataset Case: Unified Model for Fallacy Type Classification

Similar to our work in Section 4.2, previous work on fallacy recognition has tackled just one dataset at a time. For example, work on detecting propaganda techniques use fine-tuning of different pre-trained transformers with embedding-based or handcrafted features (Da San Martino et al., 2020; Jurkiewicz et al., 2020) as well as LSTMs and transformers for sequence tagging of propaganda fragments (Da San Martino et al., 2019a; Yoosuf and Yang, 2019), while Jin et al. (2022) propose a

structure-aware classifier to detect logical fallacies.

Fallacy recognition is a challenging task for three main reasons: i) the number of classification labels (fallacies types) and the class imbalance in existing datasets is often very high; ii) existing datasets cover varying genres and are typically very small in size due to annotation challenges; and iii) models trained on individual data sets often show poor out of distribution generalization. A recent line of work (Wei et al., 2022; Sanh et al., 2022) relies on the intuition that most natural language processing tasks can be described via natural language instructions and models trained on these instructions in a multitask framework show strong zero-shot performance on new tasks.

Based on this success, we propose a *unified model based on multitask instruction-based prompting* using T5 (Raffel et al., 2020) to solve the above challenges for fallacy recognition (Section 4.3.1). This approach allows us to unify all the existing datasets and a newly introduced dataset (Section 4.1.2) by converting 28 fallacy types across 5 different datasets into natural language instructions. Experimental evidence shows that our multitask fine-tuned models outperform task-specific models trained on a single dataset by an average margin of 16% as well as beat strong few-shot and zero-shot baselines by average margins of 25% and 40%, respectively in macro F1 scores across five datasets (Section 4.3.2.1). To further deepen our understanding of the task of fallacy recognition, we analyze the performance of our models for each fallacy type across datasets, model size, and prompt choice (Section 4.3.3). We further analyze the effect of annotation quality on the model performance and the feasibility of complementing this approach with external knowledge (Section 4.3.4).

4.3.1 Multitask Instruction-based Prompting

Recently, Wei et al. (2022; Sanh et al. (2022) leverage the intuition that NLP tasks can be described via natural language instructions, such as "*Is the sentiment of this movie review positive or negative?*" or "*Translate 'how are you' into Chinese.*". They then take a pre-trained language model and perform *instruction tuning* — fine-tuning the model on several NLP datasets expressed via natural language instructions. Such an approach has several benefits, the most important one is being able to have a unified model for several tasks. Finally, training tasks spanning diverse datasets in a

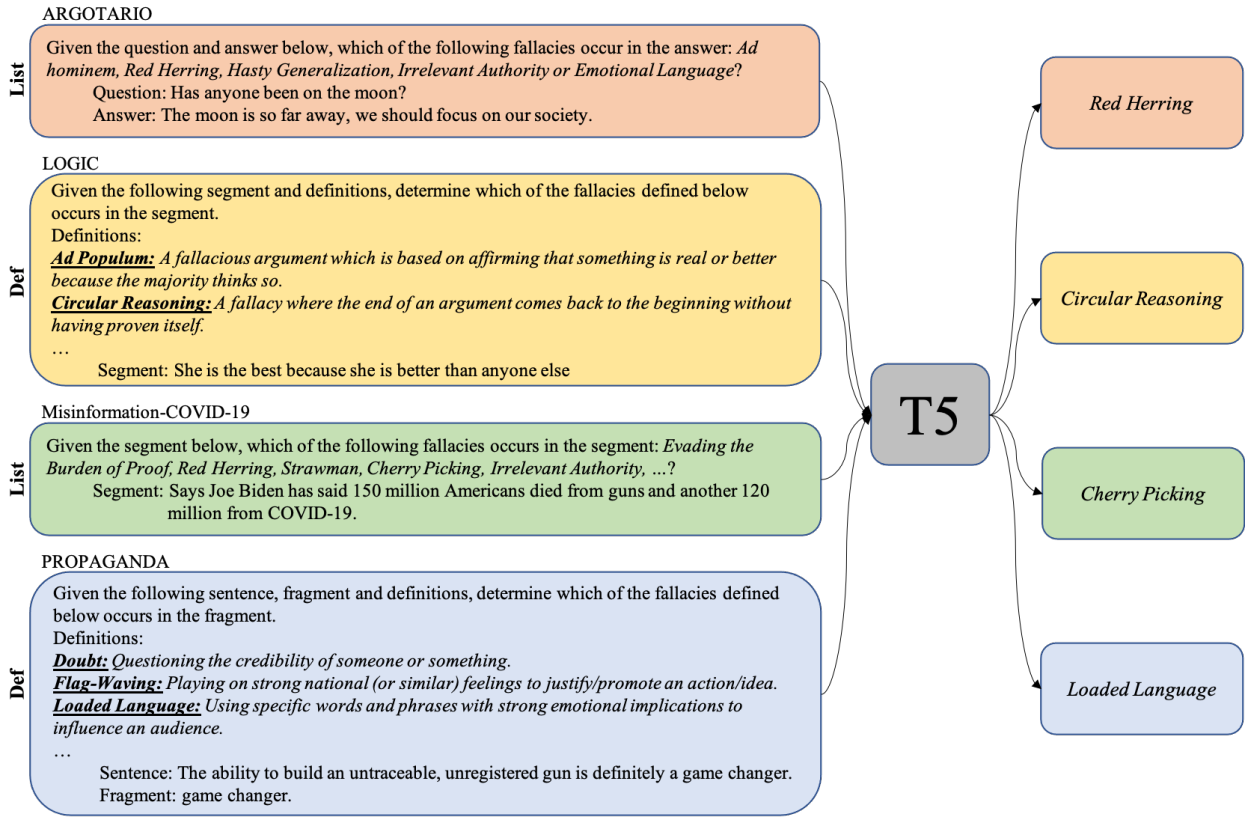


Figure 4.4: Model and Prompts. **Def**: fallacy definitions in the prompt. **List**: fallacy names listed in the prompt.

massively multitask fashion improves inference time performance, especially for smaller datasets.

Following the success of multitask instruction-based prompting, we approach different formulations of fallacies across datasets as different tasks with a generic prompting framework in a single model. We use T5 (Raffel et al., 2020) as the backbone model for training on all five fallacy datasets that have different numbers and types of fallacies. We hypothesize that when a model is able to learn to recognize fallacy from multiple datasets, it is more likely able to learn generic traits of fallacy types rather than learning characteristics specific to a single dataset.

A sample list of instructions for each dataset is shown in Figure 4.4 All instructions start with an n-gram (e.g., *'Given a text segment'*) followed by a list of fallacy types with or without their definitions. The final component of the instruction is specific to each dataset (e.g., question-answer pair for ARGOTARIO, sentence-fragment or sentence only for PROPAGANDA). The generation target during training and testing is one of the fallacy types that are permissible for each dataset. In

addition, we ask the model to generate the fragment that contains the fallacy (PROPAGANDA dataset only) during training to increase the diversity of prompts and instructions during training. Since the overall objective of this work is to have a generic classifier for fallacy and to compare it with other classification methods, evaluating the model's ability to correctly generate the fallacious fragment is beyond the scope of this paper. During inference time, we use greedy decoding and select the generated target as the prediction of fallacy type. The evaluation is done using a strict string match with the gold fallacy.

We use HuggingFace's implementation (Wolf et al., 2020) of the T5 model (large and 3B) where we train all models for 5 epochs choosing the epoch with the lowest evaluation loss as the final model. The models are run with 1e-4 learning rate, Adam optimizer, batch size 2, gradient accumulation steps 512, maximum source length 1024, and maximum target length 64. At inference time, the target is generated using greedy decoding (beam search of size 1) with no sampling and default settings for T5. The generated target is then compared with the fallacies in the given scheme and the prediction is counted as correct if they are the same using a strict string match.

4.3.2 Evaluation Setup and Results

Given the high imbalance nature of all fallacy datasets, we report both *accuracy* (equivalent to micro F1 as we do not include multi-label instances) and *Macro F1*.

Baselines. We consider the following three models as our baselines: i) zero-shot classification using UnifiedQA (Khashabi et al., 2020); ii) few-shot instruction-tuning of GPT-3 (Brown et al., 2020); and iii) full-shot fine-tuning of BERT (Devlin et al., 2019b).

UnifiedQA is a question-answering model that is trained on 20 question-answering datasets in different formats and showed generalization capability to unseen data. We use its recent version UnifiedQA-v2 (3B size) (Khashabi et al., 2022) to test the ability of a such a model to detect fallacies in zero-shot settings. The prompt for UnifiedQA follows the same format of the prompts used to train the model on the question-answering datasets. The prompt starts with the questions *Which*

ARGOTARIO	Which fallacy does the following answer to the question have: "Is television an effective tool in building the minds of children?" "Yes because the cute children are our future."? (A) Hasty Generalization (B) Red Herring (C) Emotional Language (D) Irrelevant Authority (E) Ad Hominem
PROPAGANDA	Which fallacy does the following sentence have: "But real journalism should be able to get through the shocking and the surreal and get to the truth."? (A) Irrelevant Authority (B) Red Herring (C) Causal Oversimplification (D) Name Calling or Labeling (E) Black-and-White Fallacy (F) Slogans (G) Exaggeration or Minimisation (H) Thought-terminating Cliches (I) Doubt (J) Whataboutism (K) Flag-Waving (L) Loaded Language (M) Fear or Prejudice (N) Reductio ad hitlerum

Table 4.9: Examples of for zero-shot prompts for UnifiedQA. The first example is from the ARGOTARIO dataset, which has an *emotional language* fallacy. The second example is from the PROPAGANDA dataset, which has a *loaded language* fallacy.

fallacy does the following sentence have:"sentence"? followed multiple choices of the fallacies that exist in each dataset. Table 4.9 shows two examples of the zero-shot prompts we use to test UnifiedQA. The order of the fallacies in the multiple-choice questions is shuffled for each instance to avoid learning any patterns related to how fallacies are ordered in the question.

We do few-shot instruction-tuning of GPT-3 as many fallacy datasets are of small size, which poses the need for models that can perform well using few-shot training. We set up the instructions in a similar fashion to the ones used for T5 (i.e., *List* prompt in Figure 4.4). Additionally, we set up instructions with explanations where each few-shot example has a text segment, a fallacy label, and a sentence explaining why the fallacy label is suitable for the text, which is shown to improve the results of few-shot learning (Lampinen et al., 2022). Constrained by the length allowed in the prompt, we use two-shots per the five fallacy types for the ARGOTARIO dataset and one-shot per the nine-to-fourteen fallacy types for the other datasets. Given the high number of fallacy types, it is not feasible to instruction-tune GPT-3 on the 28 unique fallacy types that exist in all five datasets combined. We use the completion API of GPT-3 from OpenAI (Brown et al., 2020) using their large engine that is trained with instructions (*text-davinci-002*) with temperature 0, max generated tokens 150 and other parameters kept at the default value (e.g., top_p=1). The generated target is considered correct if it has the gold fallacy (even with additional text). Since GPT-3 is trained with

few shots only, it sometimes generates some generic prefix, repeats the text segment, or generates more than one fallacy. We do a few-shot instruction-tuning of GPT-3 with and without explanations. The instructions that do not include explanations follow the same format of the ones shown in Figure 4.4 where it starts "Given a text segment ..." followed by a list of fallacy types and then the few-shot examples that include a text segment and a fallacy type. Additionally, we write explanations after each few-shot example in the instruction prompt, which explains why a given text segment is labeled with the fallacy type. The explanations follow the fallacy type labels as shown in Table 4.10.

For BERT, we fine-tune it for 3 epochs on each dataset separately to test its ability to do fallacy recognition. We use Huggingface's implementation of BERT (base-uncased) and fine-tune the model for 3 epochs with a 1e-5 learning rate, batch size 16, and maximum sequence length of 256.

Finally, we use a T5-large model trained on each dataset separately using the instructions shown in Figure 4.4 as a baseline to compare with the multitask setup for the same model.

4.3.2.1 Multitask Instruction-based Prompting vs. Baselines

Baseline Results Looking at the results shown in Table 4.11, UnifiedQA struggles to have any meaningful results and mostly predicts one or two fallacy types for all examples, which shows the infeasibility for models to perform well in zero-shot settings on a complex task such as fallacy recognition. GPT-3 is able to perform well on ARGOTARIO, even when trained with one-shot per class, but struggles to beat any full-shot model on the other datasets, which highlights the difficulty of this task for few-shot training. Adding the explanations does not improve the performance, which could have been outweighed by the low number of shots per class and a high number of fallacy classes. We notice that BERT has an acceptable performance on the ARGOTARIO dataset (Acc. 44% and F1 38%) that has the lowest number of classes (5 fallacy types), which is also the most balanced dataset compared to the other ones. However, when the number of fallacy classes increases to 9 or more, BERT struggles to have a good performance in any of the two evaluation measures.

The T5-large model is also trained on each dataset separately using the instructions shown in Figure 4.4. It has a surprisingly low performance on the ARGOTARIO dataset (Acc. 25% and

Given the question and answer pairs below, which of the following fallacies occur in the answers: *Emotional Language, Red Herring, Hasty Generalization, Ad Hominem*, or *Irrelevant Authority*?

———————————————————-

1) **Question:** Is Christianity a peaceful religion?
Answer: You are the antichrist, you want to destroy our belief in god.
Fallacy: Ad Hominem
Explanation: It is an ad hominem because the speaker is attacked for his bad intentions and not for the point she is making.

2) **Question:** Is television an effective tool in building the minds of children?
Answer: All TV-Shows are bad. Look at "the bachelor". Children cannot learn from it.
Fallacy: Hasty Generalization
Explanation: It is a hasty generalization since the evaluation of a whole category is drawn from the evaluation of a single element of the category.

...

5) **Question:** Should we allow animal testing for medical purposes?
Answer: No, animals are so cuuuuteeeeeeee!!!
Fallacy: Emotional Language
Explanation: It is a fallacy of emotional language since the argument appels to positive emotions associated to animals' appearances.

6) Question: Should gorillas be held in zoos
Answer: No, I don't like gorillas.

———————————————————-

Fallacy: Red Herring

Table 4.10: Example of GPT-3 few-shot instruction with explanations. The instruction ends with a Test Example that is followed by the model output containing the Generated Fallacy Type .

F1 14%) that is significantly lower than BERT and GPT-3. However, it is able to learn better for datasets with a high number of classes (13-14 classes) and large training data (e.g., PROPAGANDA and LOGIC).

Multitask Instruction-based Prompting Results We train two sizes of the T5 models (large and 3B) on all datasets combined using the instructions mentioned in Figure 4.4. This increases the performance significantly on all datasets of the T5-large model compared to its performance when trained on one dataset at a time as shown in Table 4.11. The numbers further improve when we

Training Data	Shot	Model	Argotario		Propaganda		Logic		Covid-19		Climate	
			Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
–	Zero	UnifiedQA	23	14	04	01	21	08	14	07	08	02
Single	Few	GPT-3	45	39	19	13	20	22	14	09	11	04
	Few+Exp	GPT-3	47	39	13	10	19	22	10	08	10	03
	Full	BERT	44	38	50	25	35	31	25	08	23	04
	Full	T5-Large	25	14	66	30	56	45	26	09	23	04
Multi	Full	T5-Large	59	59	70	41	68	62	**31**	26	**27**	17
	Full	T5-3B	**64**	**64**	**73**	**56**	**70**	**66**	29	**28**	25	**20**

Table 4.11: Accuracy and macro F1 scores on all datasets. **Exp**: explanations added to the few-shot examples. Numbers in **Bold** represent the best score for each dataset, and underlined numbers are the second best.

increase the size of the model from T5-large to T5-3B. This shows the benefit of our unified model based on multitask instruction-based prompting (multi-dataset) for fallacy recognition where we have limited resources and some very small datasets, and also shows the ability of larger models to generalize to the five test sets. The two multi-dataset models always have the best or second-best results on all datasets. Also, the T5-3B model is better than T5-large in all accuracy and F1 scores for all datasets except accuracy scores for the COVID-19 and CLIMATE, where the T5-large is better, which could be due to having more correct predictions in the majority classes as the T5-3B is still better in macro F1 scores. To further understand the effect of the model size and prompt choice, we discuss in the next section the per-class performance of four different T5 models.

4.3.3 Performance on Fallacy Types

We show the per-class (fallacy type) results of our unified model (multitask instruction-based prompting) using two model sizes (T5-large and T5-3B) and three prompt choices (Def, List, and All) in Tables 4.12-a to 4.12-e.

Model Size In general, increasing the model size (from T5-large to T5-3B both trained on all prompts) improves the overall results (especially macro F1) on all datasets. We notice the importance of model size in most datasets for fallacies types that have diversion moves (e.g., *Red Herring* in all datasets, *Strawman* in COVID-19 and CLIMATE, *Whataboutism* in PROPAGANDA) where additional context is usually needed to make accurate predictions. A model with more parameters is

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Black-and-White Fallacy	35	32	21	28
Causal Oversimplification	24	48	24	24
Doubt	66	69	61	60
Exaggeration or Minimization	51	61	42	37
Fear or Prejudice	44	56	45	44
Flag-Waving	58	71	73	66
Irrelevant Authority	52	49	26	36
Loaded Language	82	82	80	79
Name Calling or Labeling	83	83	82	82
Red Herring	0	50	18	0
Reductio Ad Hitlerum	0	37	0	0
Slogans	50	51	42	48
Thought-Terminating Cliches	29	44	38	24
Whataboutism	0	44	43	17
Accuracy	70	73	69	67
Macro F1	41	56	43	39

(a) Propaganda

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Ad Hominem	82	89	84	80
Ad Populum	82	86	83	80
Black-and-White Fallacy	88	84	87	89
Causal Oversimplification	70	81	65	79
Circular Reasoning	59	77	73	71
Deductive Fallacy	53	53	42	46
Emotional Language	71	68	60	57
Equivocation	29	29	29	12
Fallacy of Extension	55	51	62	18
Hasty Generalization	74	70	69	68
Intentional Fallacy	26	33	24	12
Irrelevant Authority	60	70	66	58
Red Herring	60	61	56	47
Accuracy	68	70	67	63
Macro F1	62	66	62	55

(b) Logic

Table 4.12: F1 scores for each fallacy type for two T5 model sizes (T5-Large and T5-3Billion), and for three prompt choices (**Def**: fallacy definitions in prompt; **List**: fallacy types listed in prompt; **All**: both Def and List prompts) to study the effect of model size and prompt choice. All models are trained on all five datasets combined.

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Ad Hominem	68	68	68	59
Emotional Language	67	68	68	67
Hasty Generalization	41	58	45	54
Irrelevant Authority	75	77	78	71
Red Herring	44	52	41	48
Accuracy	59	64	60	59
Macro F1	59	64	60	59

(c) Argotario

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Causal Oversimplification	29	50	44	42
Cherry Picking	31	31	35	36
Evading the Burden of Proof	47	36	27	41
False Analogy	40	33	50	33
Hasty Generalization	21	0	19	19
Irrelevant Authority	57	24	0	15
Red Herring	0	19	0	0
Strawman	0	24	0	0
Vagueness	8	31	21	0
Accuracy	31	29	28	29
Macro F1	26	28	22	21

(d) Covid-19

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Causal Oversimplification	37	29	53	32
Cherry Picking	39	41	43	41
Evading the Burden of Proof	0	0	0	0
False Analogy	25	0	0	25
Hasty Generalization	0	0	0	0
Irrelevant Authority	30	27	25	25
Red Herring	0	6	12	11
Strawman	0	46	0	25
Vagueness	22	34	26	34
Accuracy	27	25	29	28
Macro F1	17	20	18	21

(e) Climate

Table 4.13: F1 scores for each fallacy type for two T5 model sizes (T5-Large and T5-3Billion), and for three prompt choices (**Def**: fallacy definitions in prompt; **List**: fallacy types listed in prompt; **All**: both Def and List prompts) to study the effect of model size and prompt choice. All models are trained on all five datasets combined.

in principle better at capturing more information during pretraining, which could be more useful for such fallacies that require more information beyond the provided segment. This however is not always true where for other fallacies of diversion the results are the same or marginally different for the two model sizes when the fallacy is among the majority training classes (e.g., *Cherry Picking* in COVID-19 and CLIMATE), or inconsistent due to different conceptualizations of a single fallacy across datasets (e.g., *Irrelevant Authority*, more discussion at the end of this Section). Interestingly, the smaller size model (T5-large) has similar performance to the larger model (T5-3B) on some fallacy types with strong lexical cues contained in the text segment (e.g., *Loaded Language*, *Name Calling* and *Slogans* in PROPAGANDA; *Ad Hominem* and *Emotional Language* in LOGIC and ARGOTARIO).

Prompt Choice We also fix the model size (T5-3B) but change the prompts used for training to see which prompt is more useful for this task. We mainly experiment with two prompts that include either the definitions of all fallacies or only listing the names of all fallacies. In both cases, the prompt starts with an instruction followed by either definitions or fallacy names and ends with the segment that has the fallacious text. Including both prompts for each training instance yields the best results in most cases as we would expect. However, it seems that some fallacies benefit more from including the definitions in the prompt than others. In general, including the definitions (T5-3B-Def) rather than just fallacy names (T5-3B-List) has higher accuracy and macro F1 scores in 4 out of 5 datasets as shown in Table 4.12 (exceptions are accuracy in COVID-19 and F1 in CLIMATE). In particular, it seems that definitions are more useful for fallacies that are closely related to other fallacies in one scheme where the definition helps in further clarifying the difference between the two. For example, in PROPAGANDA (Table 4.12-a) *Thought-Terminating Cliches* are defined as "*words or phrases that offer short, simple and generic solutions to problems*" which is mostly confused with *Loaded Language* by most models, especially ones not trained with definitions. Also in PROPAGANDA, T5-3B-Def has a much higher score than T5-3B-List on *Whataboutism*, which is "*a discrediting technique that accuses others of hypocrisy*", which includes introducing questions

about other irrelevant matters. This could have caused models to confuse it with the *Doubt* fallacy.

Fallacy Types Across Datasets There are two fallacies that exist in all five datasets (i.e., *Irrelevant Authority* and *Red Herring*) and two other fallacies that exist in four datasets (i.e., *Causal Oversimplification* and *Hasty Generalization*). We closely look at these fallacies to understand the challenges posed by changes in the domain, genre, and annotation guidelines.

Consider the results shown in Tables 4.12 (a-e) for *Irrelevant Authority*, where we can make three observations: i) T5-large is the best in PROPAGANDA, COVID-19, and CLIMATE; ii) T5-3B-All is the best in LOGIC and marginally second best (to T5-3B-Def) in ARGOTARIO; iii) similar to model size, including the definition in the prompt has inconclusive benefit across datasets. This can be mainly attributed to inconsistency in how this fallacy is defined in different schemes as for example it strictly refers to "*mention of false authority on a given matter*" in COVID-19, while it additionally includes "*referral to a valid authority but without supporting evidence*" in PROPAGANDA. Similarly, no single model is consistently better in detecting *Red Herring* across all datasets as shown in Tables 4.12 (a-e). This, however, is more likely caused by the different format this particular fallacy has in different domains and genres as it consists of shorter phrases in PROPAGANDA, asking irrelevant or misleading questions in CLIMATE, and mentions of irrelevant entities in LOGIC.

Causal Oversimplification has more consistent results as shown in Tables 4.12 (a,b,d,e) where the T5-3B-All model has the best results in three out of four datasets. This illustrates that while the notion of this fallacy might differ across datasets, it still strongly shares common generic features (e.g., the existence of a causal relation) that make it distinguishable by a single model in different settings. Finally, the results for *Hasty Generalization* shown in Tables 4.12 (b-e) indicate that detecting this fallacy becomes more challenging when other similar fallacies exist in a fallacy scheme (e.g., *Cherry Picking* in COVID-19 and CLIMATE), and less challenging when other fallacies in the scheme are further away (e.g., LOGIC and ARGOTARIO).

Nevertheless, this multitask setup provides the model with the opportunity to learn to detect specific fallacy types as they are expressed differently, and grouped with different fallacies, which

His opinion is: "She may very well believe everything she's saying, and that is one of the signs of lunacy, believing something that isn't real." And her lawyer is even loonier Doubt Name Calling or Labeling Doubt
"Christianity is Europe's last hope," Orban told an audience of party faithful at the foot of the Royal Castle in Budapest. Slogans Flag-Waving Flag-Waving
"Orban is openly Christian and seems to understand something that many do not and that is you do not allow a wholesale flood of antichrists to pour into your country." Flag-Waving Name Calling or Labeling Flag-Waving

Table 4.14: Example sentences from PROPAGANDA with gold label , model prediction and expert annotation . Underlined text highlights the propagandistic fragment.

consistently and significantly improves the overall results of fallacy recognition over single-scheme (or single dataset) models.

4.3.4 Error Analysis

An expert looked at 70 wrongly predicted examples (5 from each of the 14 classes) from the PROPAGANDA datasets to better understand model errors and the quality of annotations. First, the expert looked only at the sentence and the fragment identified by the gold annotation as containing a fallacy and she independently annotated the propaganda technique at stake. Comparing this annotation with gold labels and model prediction (T5-3B-All), it turns out that the expert annotator agreed with the gold label in 75% of the cases, and with the model prediction in 15%, while she chose a different label in 10% of the cases. Table 4.14 shows three examples along with gold labels, model predictions, and expert annotations.

Consider the first example in Table 4.14 that has *Doubt* as the gold label. The expert agrees that the propaganda technique used rests on questioning the credibility of the lawyer (*Doubt*), even though the adjective "lunatic" is a literal instance of *Name Calling*. Thus, the label predicted by the model is not wrong but less relevant since the lack of trustworthiness is the most effective feature in undermining the antagonist's stance, regardless of whether it is due to lunacy or lack of integrity.

In the second example of Table 4.14, the expert agrees with the model prediction of a *Flag-*

Waving fallacy in the underlined segment rather than a *Slogan* as the gold label. The term "*last hope*" can be considered a slogan; however, when we consider the full propagandistic segment that includes the word "*Christianity*", it maps better to *Flag-Waving* as it has been defined in the guidelines (and included in the prompt) as "*Playing on strong national feeling (or to any group)...*". The third example highlights, even more, the importance of the selected fragment in the prompt: without considering the reference to the "*antichrist*" threat, it is not possible to understand that the sentence is playing on a religious-based national feeling.

In light of the analysis of the 70 examples in the PROPAGANDA dataset, the following general observations are found: i) some fallacious segments can map to more than one fallacy, especially when one of the two is a language fallacy (e.g., *Name Calling, Exaggeration, Loaded Language*). In such cases, the model tends to privilege the language fallacy type, even if usually not the most relevant from an argumentative perspective; ii) for some cases, the expert annotator had to read more context beyond the sentence; iii) for some cases, the expert agreed with the gold label, but disagreed with the boundaries of the annotated fragment by choosing a larger or more informative one. In light of this, improving automatic fallacy identification may entail i) considering additional context; ii) adopting a fallacy scheme with heuristics that imposes an order into fallacy recognition (structural fallacy followed by diversion and logical fallacies with language fallacies at last when all the others are excluded).

4.4 Explaining Checkworthiness Through Fallacy

Now that we have studied fallacy recognition and presented a model that is able to recognize 28 fallacies across multiple domains and genres, we are interested in using this model in explaining the reasoning behind the checkworthiness of fact-checked statements through fallacy types. As the MISINFO fallacy scheme (Section 4.1.2) was constructed by investigating statements that were fact-checked by human fact-checkers in climate change and Covid-19, it is suitable to use this fallacy scheme to explain the checkworthiness of statements in these two domains. However, as the

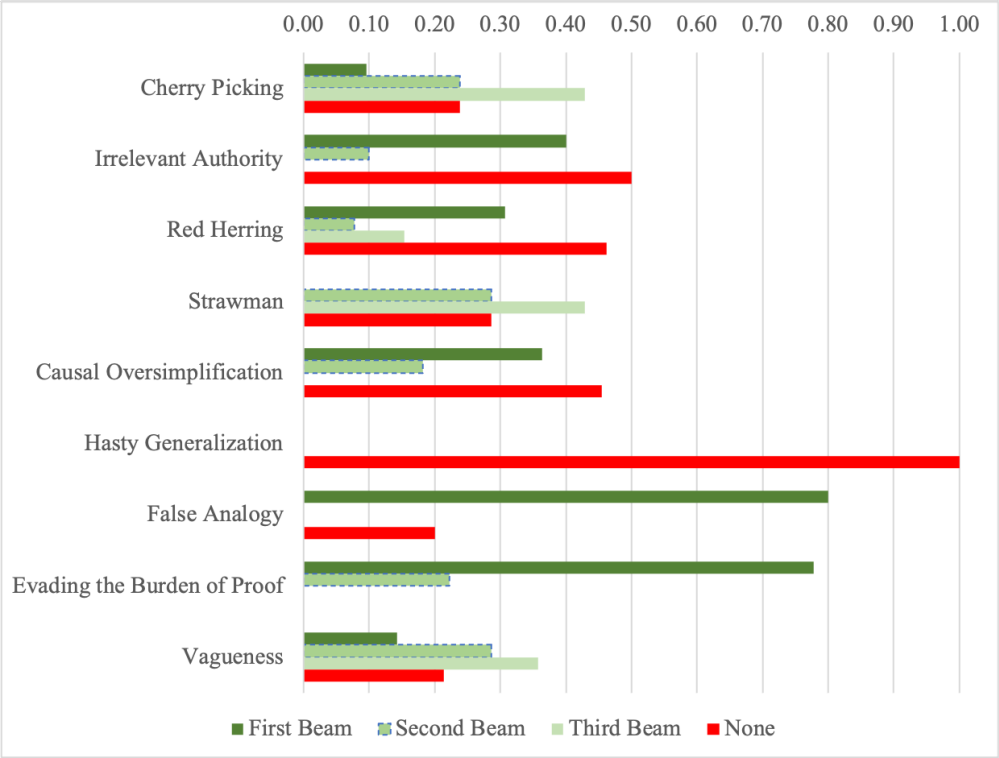


Figure 4.5: Percentage of correct (green) and wrong (red) top three beam outputs per fallacy for checkworthy statements in climate change.

gold fallacy labels show in both datasets that not all checkworthy statements have fallacies (Table 4.3), it is not expected for fallacy types to explain the checkworthiness of all statements. Therefore, we do not use fallacies for checkworthiness prediction of all statements, but rather as a way of understanding the fallacious portion of checkworthy statements.

We use our multitask fallacy recognition model on 92 and 115 fallacious checkworthy statements from the CLIMATE and COVID-19 test sets, respectively. This includes statements that are fallacious by having misleading content, false connection, or missing context and thus more likely to fall under misinformation. Non-fallacious checkworthy statements are not included in this analysis, which includes statements that are found to be true or simply presenting false information with no fallacious moves.

To generate a fallacy using our trained fallacy recognition model (T5-3B-All model in Section 4.3), we use beam search of size 3 for decoding and we return the top three predictions of fallacy for each fallacious checkworthy statement and compare them with the gold fallacy label. We show

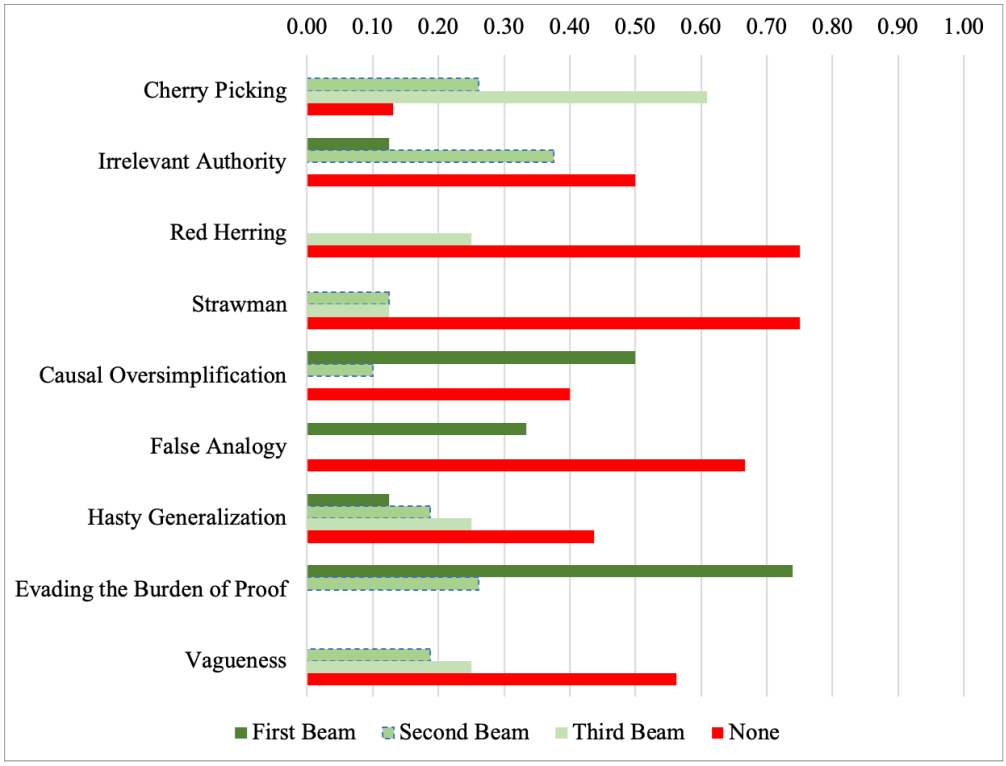


Figure 4.6: Percentage of correct (green) and wrong (red) top three beam outputs per fallacy for checkworthy statements in Covid-19.

the predicted fallacies for CLIMATE and COVID-19 in Figures 4.5 and 4.6, respectively. We see from both figures that the correct fallacy is predicted among the top three beam predictions in more than 50% of the cases for all fallacies except *Hasty Generalization* for CLIMATE, and *Red Herring, Strawman, False Analogy and Vagueness* for COVID-19. The *Irrelevant Authority* fallacy is correctly predicted among the top three beam outputs in exactly 50% of the examples and missed in the other 50%, which happens in both CLIMATE and COVID-19 datasets.

In general, the correct fallacy is among the top three beam outputs in 68% and 64% of the examples in the CLIMATE and COVID-19 datasets, respectively. This shows the possibility of using fallacies as a way to explain why a certain statement was fact-checked. The fallacy recognition model suffers from notable limitations such as i) over-prediction of some fallacy classes (e.g., *Evading the Burden of Proof*, and *Cherry Picking*) as shown in Figure 4.7; and ii) missing fallacies that might require additional context, especially in social media such as *Red Herring* and *Strawman* in the COVID-19 dataset. Although most diversion fallacies are expected to require additional

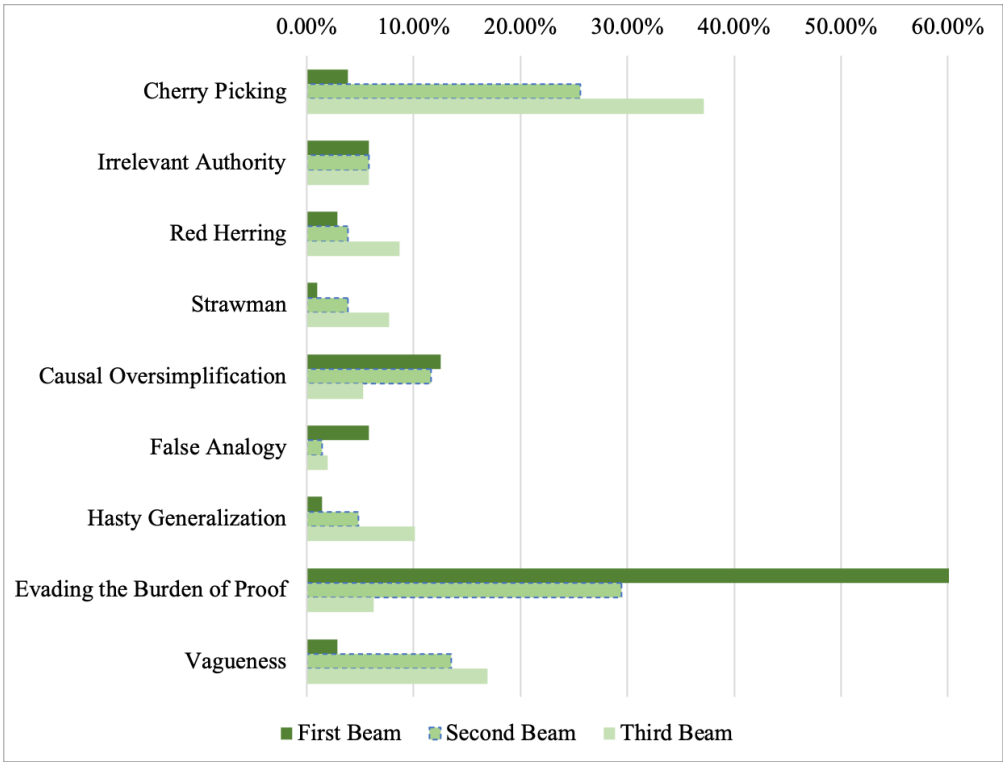


Figure 4.7: Frequencies of top three beam predictions for each fallacy type in CLIMATE and COVID-19 test sets.

context, some fallacy types (e.g., *Cherry Picking*) are detected better than others due to their higher frequency in the training data.

4.5 Conclusion

In this Chapter, we showed models for finding and classifying fallacious fragments trained on single and multiple fallacy datasets. We have also used fallacy recognition models to explain the checkworthiness of statements in the climate change and Covid-19 domains.

In a single dataset setup, we presented a sequence tagger using a BiLSTM-CRF architecture to find propaganda fragments in news articles and to classify the propaganda technique used in them. We then focused on the classification aspect of the task and fine-tuned BERT models using various input configurations that include: a fallacious fragment, a fallacious sentence, and a sentence with an annotated fallacious fragment. We show the ability of BERT to recognize fallacies when

only fallacious fragments are provided. However, the performance of BERT diminishes when fine-grained annotations are not available or when trained across multiple schemes of fallacies in a very high multi-class classification setup. Thus, presenting the need for a more unified approach to fallacy recognition.

In a multi-dataset setup, we introduced a unified model using instruction-based prompting for solving the challenges faced by the fallacy recognition task. We could unify all the datasets by converting twenty-eight fallacy types across five different datasets into natural language instructions. We showed that our unified model is better than training on a single dataset. We analyzed the effect of model size and prompt choice on the detection of specific fallacy types that could require additional knowledge better captured by bigger models (e.g., diversion fallacies such as *Red Herring*), and the distinction between similar fallacies better detected by more comprehensive prompts that include definitions of fallacy types (e.g., *Doubt* vs. *Whataboutism*). We analyzed the differences of fallacy types that appear in multiple fallacy schemes across the five datasets and showed that one fallacy type could have multiple meanings which further increases the complexity of this task (e.g., *Irrelevant Authority*). We conducted a thorough error analysis and released a new fallacy dataset for fact-checked content in the climate change domain.

Finally, we used a trained fallacy recognition model to explain the checkworthiness of fact-checked statements in two datasets as a way of using fallacies to provide reasoning for fact-checking statements. We discussed the current limitations of the model such as over-prediction of some fallacy types and missing ones that might require additional context.

Chapter 5

Verification of Statements

We have shown approaches that utilize argument structure in determining what to fact-check (Chapter 3) and argument quality through fallacy recognition that explain the reasoning behind checkworthiness (Chapter 4). We now look at the automation of the next stages of the fact-checking process by assuming we have access to a list of statements to fact-check. This requires identifying evidence from trusted sources, understanding the context, and reasoning about what can be inferred from the evidence given a target statement. Several organizations such as FactCheck.org and PolitiFact.com are devoted to such activities, and the final verdict can reflect varying degrees of truth (e.g., PolitiFact labels statements as *true, mostly true, half true, mostly false, false, and pants on fire*).

We start with end-to-end fact-checking on a binary scale where a system is given a statement to fact-check and tasked with finding relevant evidence, and judging the veracity of the statement (True/False) or deciding that the retrieved evidence is not enough and thus producing a *Not Enough Information* label. We introduce an approach to retrieve relevant evidence at the sentence level from Wikipedia as an evidence document collection and predict the veracity of the statement based on the retrieved evidence (Section 5.1). Then, we examine advancements in end-to-end automatic fact-checking systems in the literature, develop a series of adversarial attacks on evidence retrieval and claim verification, and evaluate the resilience of these systems under adversarial attacks (Section 5.2).

Next, we study the ability of machine learning models to perform fine-grained claim verification on the aforementioned six-degree truth barometer introduced by PolitiFact. We show the benefit of evidence for fine-grained claim verification by modeling the justifications provided by the human

fact-checkers (Section 5.3), which is especially crucial for naturally-occuring claims.

Our contributions in this chapter are as follows:

- We introduce an end-to-end fact-checking model that is trained and tested on the FEVER (Thorne et al., 2018a) dataset that includes claims-evidence pairs from Wikipedia. Given a claim, the model retrieves the relevant document(s) from Wikipedia, selects the most relevant evidence sentence(s), and based on the retrieved evidence, it predicts whether the claim is *Supported/True*, *Refuted/False*, or there is *Not Enough Information*.

- We enhance the FEVER dataset by introducing linguistic challenges that are common in naturally occurring claims such as multi-hop propositions, temporal reasoning, named entity ambiguity, and lexical variation. We show that state-of-the-art fact-checking systems are vulnerable to adversarial attacks from this dataset.

- We address complex claims that cannot be given veracity labels using a binary scale (True/False), and we show the importance of human justifications for fine-grained claim verification on six levels of truth.

5.1 Evidence Retrieval and Claim Verification

We developed one of the first end-to-end systems for fact extraction and verification that was ranked 6th (out of 24) in the first Fact Extraction and VERification (FEVER) shared task in 2018 (Chakrabarty et al., 2018; Thorne et al., 2018c). Our fact-checking system is tasked with both evidence retrieval and claim verification. Due to the complexity of open-ended evidence retrieval and continuous change of facts, we limit the source of evidence to one dump of Wikipedia from the year 2018 as introduced by the FEVER shared task and dataset (Thorne et al., 2018a).

This dataset has enabled the development of end-to-end fact-checking systems, requiring document retrieval and evidence sentence extraction to corroborate a veracity relation prediction. The task aims to evaluate the ability of a system to verify information using evidence from Wikipedia. Given a claim involving one or more entities (mapping to Wikipedia pages), the system must extract

Claim : Fox 2000 Pictures released the film Soul Food.
[wiki/Soul_Food_(film)]
Evidence: Soul Food is a 1997 American comedy-drama film produced by Kenneth "Babyface" Edmonds , Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures .
Verdict: SUPPORTS

Claim : Murda Beatz's real name is Marshall Mathers.
[wiki/Murda_Beatz]
Evidence: Shane Lee Lindstrom (born February 11, 1994), known professionally as Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario.
Verdict: REFUTES

Claim : L.A. Reid has served as the CEO of Arista Records for four years.
[wiki/L.A._Reid]
Evidence: He has served as the chairman and CEO of Epic Records, a division of Sony Music Entertainment, the president and CEO of Arista Records, and the chairman and CEO of the Island Def Jam Music Group.
Verdict: NOT ENOUGH INFO

Figure 5.1: Examples of claims, the extracted evidence from Wikipedia and the verdicts from the FEVER dataset (Thorne et al., 2018a).

textual evidence (sets of sentences from Wikipedia pages) that supports or refutes the claim and then using this evidence, it must label the claim as Supported, Refuted or NotEnoughInfo. The FEVER dataset (Thorne et al., 2018a) was created by extracting sentences from popular Wikipedia pages and mutating them with paraphrases or other edit operations to create a claim. Then, each claim was labeled and paired with evidence or the empty set for NEI. Overall, there are 185,445 claims, of which 90,367 are S, 40,107 are R, and 45,971 are NEI. Figure 5.1 shows three instances from the data set with the claim, the evidence and the verdict.

The baseline system described by Thorne et al. (2018a) uses three major components:

- **Document Retrieval**: Given a claim, identify relevant documents from Wikipedia which contain the evidence to verify the claim. Thorne et al. (2018a) used the document retrieval component from the DrQA system (Chen et al., 2017b), which returns the k nearest documents for a query using cosine similarity between binned unigram and bigram TF-IDF vectors.

- **Sentence Selection**: Given the set of retrieved documents, identify the candidate evidence sentences. Thorne et al. (2018a) used a modified document retrieval component of DrQA (Chen et al., 2017b) to select the top most similar sentences w.r.t the claim, using bigram TF-IDF with binning.

- **Textual Entailment**: For the entailment task, training is done using labeled claims paired with evidence (labels are SUPPORTS, REFUTES, NOT ENOUGH INFO). Thorne et al. (2018a) used the decomposable attention model (Parikh et al., 2016) for this task. For the case where multiple sentences are required as evidence, the strings were concatenated.

Our system implements changes in all three modules (Section 5.1.1), which leads to significant improvements both in the development and test sets. On the shared task's development set, our document retrieval approach covers 94.4% of the claims requiring evidence, compared to 55.30% in the baseline. Further, on the development set our evidence recall is improved by 33 points over the baseline. For entailment, our model improves the baseline by 7.5 points on the development set. Overall, our end-to-end system shows an improvement of 19.56 in FEVER score compared to the baseline (50.83 vs. 31.27) on the development set. On the blind test set, we achieve an evidence recall of 75.89 and an entailment accuracy of 57.45 (9 points above baseline) resulting in a FEVER score of 49.06 (Section 5.1.2). Together with the results we discuss some lessons learned based on our error analysis.

5.1.1 Method

5.1.1.1 Document Retrieval

Document retrieval is a crucial step when building an end-to-end system for fact extraction and verification. Missing a relevant document could lead to missed evidence, while non-relevant documents would add noise for the subsequent tasks of sentence selection and textual entailment. We propose a multi-step approach for retrieving documents relevant to the claims.

- **Google Custom Search API**: Wang et al. (2018) looked at retrieving relevant documents for

fact-checking articles, looking at generating candidates via search. Inspired by this, we first use the Custom Search API of Google to retrieve documents having information about the claim. We add the token `wikipedia` to the claim and issue a query and collect the top 2 results.

- **Named Entity Recognition**: Second, we use the AllenNLP (Gardner et al., 2017) pre-trained bidirectional language model (Peters et al., 2017) for named entity recognition [1]. After finding the named entities in the claim, we use Wikipedia python API [2] to collect the top Wikipedia document returned by the API for each named entity.

- **Dependency Parse**: Third, to increase the chance of detecting relevant entities in the claim, we find the first lowercase verb phrase (VP) in the dependency parse tree and query the Wikipedia API with all the tokens before the VP. The reason for emphasizing lowercase verb phrases is to avoid missing entities in claims such as “Finding Dory was directed by X", where the relevant entity is “Finding Dory".

  To deal with entity ambiguity, we also add the token `film` in our query where the claim contains keywords such as `film`, `stars`, `premiered`, and `directed by`. For example, in Marnie was directed by Whoopi Goldberg., Marnie can refer to both wikipedia pages `Marnie (film)` and `Marnie`. Our point of interest here is `Marnie (film)`. We only experimented with `film` to capture the performance gains. One of our future goals is to build better computational models to handle entity ambiguity or entity linking.

- **Combined**: We use the union of the documents returned by the three approaches as the final set of relevant documents to be used by the sentence selection module.

Table 5.1 shows the percentage of claims that can be fully supported or refuted by the retrieved documents before sentence selection on the development set. We see that our best approach (combined) achieved a high coverage of 94.4% compared to the baseline (Thorne et al., 2018a) of

[1] http://demo.allennlp.org/named-entity-recognition
[2] https://pypi.org/project/wikipedia/

Method	Avg k	Coverage
Google API	2	79.5%
NER	2	77.1%
Dependency Parse	1	80.0%
Combined	3	**94.4%**
(Thorne et al., 2018a)	5	55.3%

Table 5.1: Coverage of claims that can be fully supported or refuted by the retrieved documents (development set).

55.3%. Because we do not have the gold evidence for the blind test set we cannot report the claim coverage using our pipeline. Our document retrieval component was the most superior among all other shared task systems at that time (2018). It was later incorporated in more recent state-of-the-art systems for fact-checking (e.g., (Hidey et al., 2020))

5.1.1.2 Sentence Selection

For sentence selection, we use the modified document retrieval component of DrQA (Chen et al., 2017b) to select sentences using bigram TF-IDF with binning as proposed by (Thorne et al., 2018a). We extract the top 5 most similar sentences from the k most relevant documents using the TF-IDF vector similarity. Our evidence recall is 78.4 as compared to 45.05 in the development set of FEVER (Thorne et al., 2018a), which demonstrates the importance of document retrieval in fact extraction and verification. On the blind test set our sentence selection approach achieves an evidence recall of 75.89.

However, even though TF-IDF proves to be a strong baseline for sentence selection, we notice on the development set that using all of the five evidence sentences together introduces additional noise to the entailment model. To solve this, we further filter the top three evidence sentences from the selected five evidence sentences using distributed semantic representations. Peters et al. (2018) show how deep contextualized word representations model both complex characteristics of word use (e.g., syntax and semantics) and usage across various linguistic contexts. Thus, we use the ELMo embeddings to convert the claim and the evidence to vectors. We then calculate the cosine similarity between the claim and the evidence vectors and extract the top three sentences based

on the score. Because there is no penalty involved for poor evidence precision, we return all five selected sentences as our predicted evidence, but use only the top three sentences for the entailment model.

5.1.1.3 Claim Verification

The final stage of our pipeline is recognizing textual entailment. Unlike Thorne et al. (2018a), we do not concatenate the evidence sentences, but train our model for each claim-evidence pair. For recognizing textual entailment, we use the model introduced by Conneau et al. (2017a) in their work on supervised learning of universal sentence representations.

The architecture is presented in Figure 1. We use bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) with max-pooling to encode the claim and the evidence. The text encoder provides a dense feature representation of an input claim or evidence. Formally, for a sequence of T words $w_{t=1,..,T}$, the BiLSTM layer generates a sequence of h_t vectors, where h_t is the concatenation of a forward and a backward LSTM output. The hidden vectors h_t are then converted into a single vector using max-pooling, which chooses the maximum value over each dimension of the hidden units. Overall, the text encoder can be treated as an operator (Text $\rightarrow R^d$) that provides d dimensional encoding for a given text.

Out-of-vocabulary issues in pre-trained word embeddings are a major bottleneck for sentence representations. To solve this, we use fastText embeddings (Bojanowski et al., 2017) which rely on subword information. Also, these embeddings were trained on a Wikipedia corpus making them an ideal choice for this task.

As shown in Figure 5.2, the shared sentence encoder outputs a representation for the claim u and the evidence v. Once the sentence vectors are generated, the following three methods are applied to extract relations between the claim and the evidence: (i) concatenation of the two representations (u, v); (ii) element-wise product u*v and (iii) absolute element-wise difference $|u - v|$. The resulting vector, which captures information from both the claim and the evidence, is fed into a 3-class classifier consisting of fully connected layers culminating in a softmax layer.

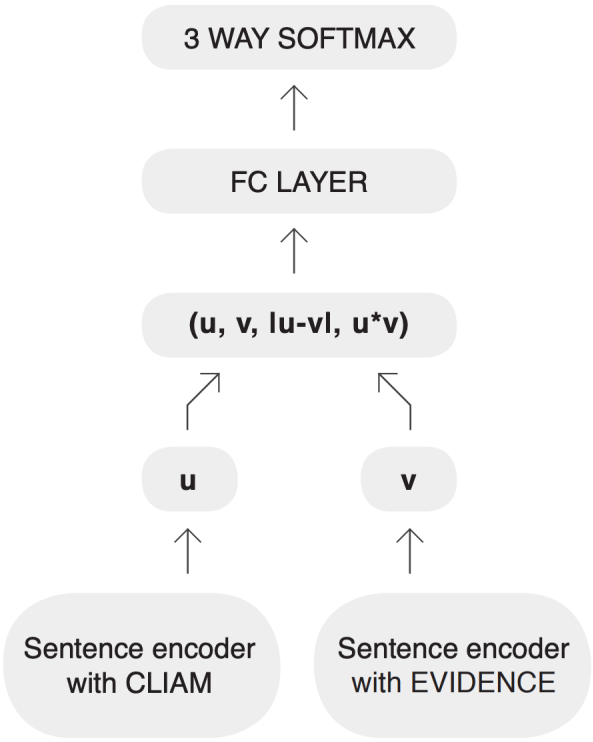


Figure 5.2: The architecture for recognizing textual entailment (Conneau et al., 2017a).

For the final class label, we experimented first with taking the majority prediction of the three (claim, evidence) pairs as our entailment label, but this led to lower accuracy on the development set. So, our final predictions are based on the rule outlined in Algorithm 1, where SUPPORTS = S, REFUTES = R, NOT ENOUGH INFO = N and C is a count function. Because the selected evidence sentences are inherently noisy and our pipeline does not concatenate them together, we choose this rule over majority prediction to mitigate the dominance of prediction of NOT ENOUGH INFO class.

$$C(S) = 1 \;\&\; C(N) = 2 \; label = S \; C(R) = 1 \;\&\; C(N) = 2 \; label = R$$
$$label = \arg\max(C(S), C(R), C(N))$$

We also experimented with training a classifier that takes confidence scores of all three claim-evidence pairs along with their positions in the document and trained a boosted tree classifier, but the accuracy did not improve. Empirically, the rule gave us the best results on the development set and thus we use it to obtain the final label.

Table 5.2 shows the three-way classification accuracy using the textual entailment model described above.

DataSet	Accuracy
Shared Task Dev	58.77
Blind Test Set	57.45

Table 5.2: Three way classification results.

DataSet	Recall
Shared Task Dev	78.4
Blind Test Set	75.89

Table 5.3: Evidence recall on development and test set.

Our entailment accuracy on the shared task development and test set is 7 and 9 points better than the baseline, respectively.

Implementation Details. The batch size is kept at 64. The model is trained for 15 epochs using the Adam optimizer with a learning rate of 0.001. The size of the LSTM hidden units is set to 512 and for the classifier, we use an MLP with one hidden layer of 512 hidden units. The embedding dimension of the words is set to 300.

5.1.2 End-to-End Results and Error Analysis

Table 5.4 shows the overall FEVER score obtained by our pipeline on the development set and on the test set. In the provisional ranking, our system is ranked sixth.

Data	Pipeline	FEVER
DEV	(Thorne et al., 2018a)	31.27
	Ours	**50.83**
TEST	(Thorne et al., 2018a)	27.45
	Ours	**49.06**

Table 5.4: FEVER scores on shared task development and test set.

On closer investigation, we find that neither TF-IDF nor sentence-embedding-based approaches are perfect when it comes to sentence selection, although TF-IDF works better.

Fox 2000 Pictures released the film Soul Food	0.29
Soul Food is a 1997 American comedy-drama film produced by Kenneth "Babyface" Edmonds,Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures	

Table 5.5: Cosine similarity between claim and supporting evidence.

Table 5.5 goes on to prove that we cannot rely on models that entirely depend on semantics. In spite of the two sentences being similar, the cosine similarity between them is poor mostly because the evidence contains a lot of extra information which might not be relevant to the claim and can be difficult for the model to understand.

At seventeen or eighteen years of age, he joined Plato's Academy in Athens and remained there until the age of thirty-seven (c. 347 BC)
Shortly after Plato died, Aristotle left Athens and at the request of Philip II of Macedon, tutored Alexander the Great beginning in 343 BC

Table 5.6: The top evidence is selected by annotators and the bottom evidence by our pipeline.

We also found instances where the predicted evidence is correct, but it does not match the gold evidence. For the claim "Aristotle spent time in Athens", both pieces of evidence given in Table 5.6 support it, but still our system gets penalized for not being able to match the gold evidence.

We found quite a few annotations to be incorrect and hence the FEVER scores are lower than expected. Table 5.7 shows two instances where the gold labels for the claims are NOT ENOUGH INFO, while in fact, they should have been SUPPORTS and REFUTES, respectively.

Table 5.8 reflects the fact that NOT ENOUGH INFO is often hard to predict and that is where our model needs to improve more.

The lines between SUPPORTS and NOT ENOUGH INFO are often blurred as shown in Table 5.8.

Claim: Natural Born Killers was directed by Oliver Stone **Evidence**: Natural Born Killers is a 1994 American satirical crime film directed by Oliver Stone and starring Woody Harrelson , Juliette Lewis , Robert Downey Jr. , Tom Sizemore , and Tommy Lee Jones .
Claim:Anne Rice was born in New Jersey **Evidence**: Born in New Orleans, Rice spent much of her early life there before moving to Texas, and later to San Francisco

Table 5.7: Wrong gold label (NOT ENOUGH INFO).

	S	N	R
S	**4635**	1345	686
N	2211	**3269**	1186
R	1348	1470	**3848**

Table 5.8: Confusion matrix of entailment predictions on the shared task development set.

Our models need a better understanding of semantics to be able to identify these. Table 5.9 shows one such example where the `gospel` keyword becomes the discriminative factor.

Claim: Happiness in Slavery is a gospel song by Nine Inch Nails
Evidence: Happiness in Slavery, is a song by American industrial rock band Nine Inch Nails from their debut extended play (EP), Broken(1992)

Table 5.9: Example where our model predicts SUPPORTS for a claim labeled as NOT ENOUGH INFO.

5.2 Fact-Checking Systems Under Adversarial Attacks

Since the claims in FEVER were manually written using information from Wikipedia, the dataset may lack linguistic challenges that occur in verifying naturally occurring checkworthy claims, such as temporal reasoning or lexical generalization/specification. Thorne and Vlachos (2019) designed a second shared task (FEVER 2.0) for participants to create adversarial claims ("attacks") to break state-of-the-art systems and then develop systems to overcome those attacks.

We present a novel dataset of adversarial examples for fact extraction and verification in three challenging categories: 1) multiple propositions (claims that require multi-hop document or sentence retrieval); 2) temporal reasoning (date comparisons, ordering of events); and 3) named entity ambiguity and lexical variation. We show that state-of-the-art systems are vulnerable to adversarial attacks from this dataset.

5.2.1 Advancement in Automating Fact-Checking

Below we describe the seven fact-checking models we selected for adversarial attacks.

Baseline (Thorne et al., 2018a): Document Retrieval using DrQA system (Chen et al., 2017a), which returns the k nearest documents for a query using cosine similarity between binned unigram and TF-IDF vectors. Sentence selection ranks sentences by TF-IDF similarity to the claim. RTE is done using Parikh et al. (2016)'s model with decomposable attention for entailment.

Papelo (Malon, 2018): develop a high precision entailment classifier based on transformer networks pretrained with language modeling (Radford et al., 2018), to classify a broad set of potential evidence. They include the articles best matching the claim text by TF-IDF score, and additional articles whose titles match named entities and capitalized expressions occurring in the claim text. The entailment module evaluates potential evidence one statement at a time, together with the title of the page the evidence came from.

Athene (Hanselowski et al., 2018b): Apply the constituency parser from AllenNLP to extract noun phrases in the claim and make use of Wikipedia API to search corresponding pages for each noun phrase, and stemmed the words of their titles and the claim, and then discarded pages whose stemmed words of the title are not completely included in the set of stemmed words in the claim. For sentence selection, the hinge loss with negative sampling is applied to train the enhanced LSTM. For a given positive claim-evidence pair, negative samples are generated by randomly sampling sentences from the retrieved documents. For RTE, they combine the five sentences from sentence selection and the claim to form five pairs and then apply enhanced LSTM for each pair. They combine the resulting representations using average and max pooling and feed the resulting vector

through an MLP for classification.

UCL (Yoneda et al., 2018): Document retrieval attempts to find the name of a Wikipedia article in the claim, and then ranks each article based on capitalization, sentence position, and token match features. A set of sentences are then retrieved from the top-ranked articles, based on token matches with the claim and position in the article. A natural language inference model is then applied to each of these sentences paired with the claim, giving a prediction for each potential evidence. These predictions are then aggregated using a simple MLP, and the sentences are reranked to keep only the evidence consistent with the final prediction.

UNC (Nie et al., 2019a): The document retriever chooses candidate wiki documents via matching of keywords between the claims and the wiki-document titles, also using external pageview frequency statistics for wiki-page ranking. The sentence selector is a sequence-matching neural network that conducts a further fine-grained selection of evidential sentences by comparing the given claim with all the sentences in the candidate documents. This module is trained as a binary classifier that is given the ground truth evidence as positive examples and all the other sentences as negative examples with an annealing sampling strategy. Finally, the claim verifier (with WordNet and ELMo features) takes the concatenation of all selected evidence as the premise and the claim as the hypothesis, and labels each such evidences-claim pair as one of 'support', 'refute', or 'not enough info'. Also, feeding the sentence similarity score (produced by the sentence selector) as an additional token-level feature to the claim verifier.

Dominik (Stammbach and Neumann, 2019): propose a two-staged sentence selection strategy to account for examples in the dataset where evidence is not only conditioned on the claim but also on previously retrieved evidence. They use a publicly available document retrieval module (Athene's (Hanselowski et al., 2018b)) and have fine-tuned BERT checkpoints for sentence selection and as the entailment classifier.

Ours 2.0 (Hidey et al., 2020): We perform document ranking by selecting the top $D < M$ pages with a pointer network. In order to obtain representations as input to the pointer network

for document ranking, we leverage the fact that Wikipedia articles all have a title (e.g., *[Barack Obama]*), and fine-tune BERT on title and claim pairs, in lieu of examining the entire document text (which due to its length is not suitable for BERT). Because the title often overlaps lexically with the claim (e.g., *[Michelle Obama]*), we can train the model to locate the title in the claim. Furthermore, the words in the title co-occur with words in the article (e.g., *Barack* and *Michelle*), which the pre-trained BERT language model may be attuned to. We thus fine-tune a classifier on a dataset created from title and claim pairs (where positive examples are titles of gold evidence pages and negative are randomly sampled from our candidate set), obtaining 90.0% accuracy.

The sentence selection and relation prediction tasks are closely linked, as predicting the correct evidence is necessary for predicting *Support* and *Attack* and the representation should reflect the interaction between a claim and evidence set. Conversely, if a claim and evidence set are unrelated, the model should predict NEI. We thus jointly model this interaction by sharing the parameters of the pointer network - the hidden state of the decoder is used for both tasks and the models differ only by a final MLP. In this model, we fine-tune a classifier on claim and evidence sentence pairs to obtain BERT embeddings on veracity relation prediction. We create a dataset by pairing each claim with its set of gold evidence sentences.

As gold evidence is not available for NEI relations, we sample sentences from our candidate documents to maintain a balanced dataset. We then fine-tune a BERT classifier on relation prediction, obtaining 93% accuracy. In order to closely link veracity relation prediction with evidence prediction, we re-frame the task as a sequence labeling task. In other words, rather than making a single prediction given all evidence sentences, we make one prediction at every time-step during decoding to model the relationship between the claim and *all evidence retrieved to that point*. This approach provides three benefits: it allows the model to better handle noise (when an incorrect evidence sentence is predicted), to handle multi-hop inference (to model the occurrence of switching from NEI to S/R), and to effectively provide more training data (for $k = 5$ timesteps we have five times as many relation labels).

5.2.2 Adversarial Dataset for Fact-Checking

We describe below the three types of attacks we introduce by performing alterations on the FEVER dataset: 1) multiple propositions (claims that require multi-hop document or sentence retrieval); 2) temporal reasoning (date comparisons, ordering of events); and 3) named entity ambiguity and lexical variation. In the discussion below, we refer to the three veracity labels as S for *Support*, R for *Refute*, and NEI for *Not Enough Info*.

5.2.2.1 Multiple Propositions

Checkworthy claims often consist of multiple propositions (Graves, 2018). In the FEVER task, checking these claims may require retrieving evidence sequentially after resolving entities and events, understanding discourse connectives, and evaluating each proposition.

Consider the claim *"Janet Leigh was from New York and was an author."*, the Wikipedia page **[Janet Leigh]** contains evidence that she was an author, but makes no mention of New York. We generate new claims of the CONJUNCTION type *automatically* by mining claims from FEVER and extracting entities from the subject position. We then combine two claims by replacing the subject in one sentence with a discourse connective such as "and." The new label is S if both original claims are S, R if at least one claim is R, and NEI otherwise.

While CONJUNCTION claims provide a way to evaluate multiple propositions about a single entity, these claims only require evidence from a single page; hence we create new examples requiring reasoning over multiple pages. To create MULTI-HOP examples, we select claims from FEVER whose evidence obtained from a single page P contains at least one other entity having a valid page Q. We then modify the claim by appending information about the entity, which can be verified from Q. For example, given the claim *"The Nice Guys is a 2016 action comedy film."*, we make a multi-hop claim by obtaining the page **[Shane Black]** (the director) and appending the phrase *"directed by a Danish screenwriter known for the film Lethal Weapon."*

While multi-hop retrieval provides a way to evaluate the S and R cases, composition of multiple propositions may also be necessary for NEI, as the relation between the claim and evidence

may be changed by more general/specific phrases. We thus add ADDITIONAL UNVERIFIABLE PROPOSITIONS that change the gold label to NEI. We selected claims from FEVER and added propositions which have no evidence in Wikipedia (e.g., for the claim *"Duff McKagan is an American citizen,"* we can add the reduced relative clause *"born in Seattle"*).

5.2.2.2 Temporal Reasoning

Many checkworthy claims contain dates or time periods and verifying them requires models that can handle temporal reasoning (Thorne and Vlachos, 2017).

In order to evaluate the ability of current systems to handle temporal reasoning, we modify claims from FEVER. More specifically, using claims with the phrase "in <date>" we *automatically* generate seven modified claims using simple DATE MANIPULATION heuristics: arithmetic (e.g., *"in 2001" → "4 years before 2005"*), range (*"in 2001" → "before 2008"*), and verbalization (*"in 2001" → "in the first decade of the 21st century"*).

We also create examples requiring MULTI-HOP TEMPORAL REASONING, where the system must evaluate an event in relation to another. Consider the S claim *"The first governor of the Indiana Territory lived long enough to see it become a state."* A system must resolve entity references (Indiana Territory and its first governor, William Henry Harrison) and compare dates of events (the admittance of Indiana in 1816 and the death of Harrison in 1841). While multi-hop retrieval may resolve references, the model must understand the meaning of *"lived long enough to see"* and evaluate the comparative statement. To create claims of this type, we mine Wikipedia by selecting a page X and extracting sentences with the pattern "is/was/named the A of Y" (e.g., A is *"first governor"*) where Y links to another page. Then we manually create temporal claims by examining dates on X and Y and describing the relation between the entities and the events.

5.2.2.3 Named Entity Ambiguity and Lexical Variation

As fact-checking systems are sensitive to lexical choice (Nakashole and Mitchell, 2014; Rashkin et al., 2017a), we consider how variations in entities and words may affect veracity relation prediction.

Attack	Seed Claim	Modified Claim
Conjunction (**A**)	Janet Leigh was from New York. Janet Leigh was an author.	Janet Leigh was from New York *and* was an author.
Multi-hop	The Nice Guys is a 2016 action comedy film.	+ directed by a Danish screenwriter known for the 1987 action film Lethal Weapon.
Date Manip. (**A**)	in 2001 in 2009	in the first decade of the 21st century 3 years before 2012 *(or)* in the 2010s
Multi-hop Temp.	The first governor of the Indiana Territory died in 1841.	The first governor of the Indiana Territory lived long enough to see it become a state.
Entity Dis.	Kate Hudson is an American actress.	Kate Hudson is a left wing political activist.
Lexical Subs. (**A**)	The Last Song began filming in 2009.	The Last Song began *shooting* in 2009.

Table 5.10: Examples of adversarial attacks. (**A**: generated automatically).

ENTITY DISAMBIGUATION has been shown to be important for retrieving the correct page for an entity among multiple candidates (Hanselowski et al., 2018b). To create examples that contain ambiguous entities, we selected claims from FEVER where at least one Wikipedia disambiguation page was returned by the Wikipedia Python API.[3] We then created a new claim using one of the documents returned from the disambiguation list. For example the claim *"Patrick Stewart is someone who does acting for a living."* returns a disambiguation page, which in turn gives a list of pages such as **[Patrick Stewart]** and **[Patrick Maxwell Stewart]**.

Finally, as previous work has shown that neural models are vulnerable to LEXICAL SUBSTITUTION (Alzantot et al., 2018), we apply their genetic algorithm approach to replace words via counter-fitted embeddings. We make a claim adversarial to a model fine-tuned on claims and gold evidence by replacing synonyms, hypernyms, or hyponyms, e.g., *created* $\rightarrow$ *established*, *leader* $\rightarrow$ *chief*. We manually remove ungrammatical claims or incorrect relations.

We show examples of the seed and generated claims from all types in Table 5.10.

5.2.3 Resilience of Fact-Checking Systems

We show the performance of seven fact-checking models (described in Section 5.2.1) under the aforementioned adversarial attacks in Table 5.11. Our model has the best overall label accuracy of 48% and the second best FEVER score of 43%. The highest model in FEVER score is *Papelo*, which fine-tunes a transformer only on claims that require a single evidence and ignores all other claims in

[3]https://pypi.org/project/wikipedia/

	Conj.		MH[1]		MH[2]		DM		MH-T		ED		LS		Overall*	
System	LA	FS	LA	FS	LA	FS	LA	FS	LA	FS	LA	FS	LA	FS	LA	FS
Baseline	.27	.17	**.67**	.10	0	0	.44	.20	**.60**	**.10**	**.83**	.54	.51	.21	.41	.17
Athene	.41	.41	.34	0	0	0	.23	.21	.07	0	.37	.26	**.83**	**.75**	.37	.29
UNC	.31	.31	.24	0	.78	.78	.27	.27	.17	0	.75	.67	.40	.36	.43	.38
UCL	.58	.58	.26	0	0	0	.47	.39	.07	0	.71	.63	.46	.35	**.48**	.39
Dominik	.37	.37	.50	**.28**	0	0	.29	.27	.13	.03	.79	**.75**	.48	.45	.47	.43
Ours	.41	.41	.34	.17	0	0	**.55**	.42	.03	0	.79	.71	.30	.26	**.48**	.43
Papelo	**.66**	**.66**	.10	0	**.94**	**.94**	.43	**.43**	.03	.03	.38	.50	.27	.33	.47	**.45**

Table 5.11: Performance of seven fact-checking models under adversarial attacks ordered by overall FEVER score. **MH[1]**: Multi-hop (S,R) labels, **MH[2]**: Multi-hop (NEI) label, **DM**: Date Manipulation, **MH-T**: Multi-hop temporal reasonsing, **ED**: Entity Disambiguation, **LS**: Lexical substitution. Evaluation metrics: **LA**: Label Accuracy, **FS**: FEVER Score. *Attack counts are not equal across types and include other adversarial attacks not shown here.

the training data requiring multiple evidence. They start with an (NEI) label for all instances and only change it to S, R upon retrieving relevant evidence. Thus, it is not surprising that the model's highest performance is on multi-hop attacks that are labeled as NEI (MH^2), and the lowest on the ones labeled S, R (MH^1) as it is not doing any multi-hop handling.

Similarly, artificially high scores are found in the baseline model by Thorne et al. (2018a), where it has the highest label accuracy on both MH^1 and MH-T. However, the model has very low FEVER score (that is a combination of accuracy and evidence recall), which indicates that the model is not retrieving the correct evidence and simply guessing the labels based on the more frequent S label.

The main challenge by multi-hop attacks is in evidence retrieval, therefore high label accuracy is not enough on these attacks. In our model, we add a post-processing step to handle temporal claims. This results in having the highest label accuracy on the date manipulation attack. However, it seems that combining temporal and multi-hop reasoning results in the most challenging attack (MH-T) to handle for all systems.

5.3 Human Justifications for Fine-Grained Claim Verification

We have introduced models for end-to-end fact-checking where the statements can be given a True (supported)/False (Refuted) veracity label on a binary scale, or deciding there is Not Enough Information. However, some statements could be true in some contexts or time-frames but false in

other situations. Thus, we study the problem of fine-grained claim verification in this section.

Wang (2017) has introduced a large dataset (LIAR) of claims from POLITIFACT, the associated metadata for each claim, and the verdict (6 class labels). Most work on the LIAR dataset has focused on modeling the content of the claim (including hedging, sentiment, and emotion analysis) and the speaker-related metadata (Wang, 2017; Rashkin et al., 2017a; Long et al., 2017). However, these approaches do not use the evidence and the justification provided by humans to predict the label. Extracting evidence from (trusted) sources for fact-checking or for argument mining is a difficult task (Rinott et al., 2015; Thorne et al., 2018a; Baly et al., 2018b). Initial fact-checking approaches rely on the fact-checking article associated with the claim. We extend the original LIAR dataset by automatically extracting the justification given by humans for labeling the claim, from the fact-checking article (Section 5.3.1) (Alhindi et al., 2018). We release the extended LIAR dataset (LIAR-PLUS) to the community.[4]

We show that modeling the extracted justification in conjunction with the claim (and metadata) provides a significant improvement regardless of the machine learning model used (feature-based or deep learning) both in a binary classification task *(true, false)* and in a six-way classification task *(pants on fire, false, mostly false, half-true, mostly true, true)* (Section 5.3.3). We provide a detailed error analysis and per-class results. This work was done in 2018 (Alhindi et al., 2018) and it focuses on feature-based machine learning models and some of the earlier deep learning models, but more recent deep learning models such as DeBERTa (He et al., 2021) could also be applied.

Our work complements the other work on providing datasets and models that enable the development of an end-to-end pipeline for fact-checking such as work by Thorne et al. (2018a) for English and Baly et al. (2018b) for Arabic. We are primarily concerned with showing the impact of modeling the human-provided justification for predicting the veracity of a claim on a six-level scale of truthfulness. We aim in this setup to capture the varying degrees of truth that some claims might have and that is usually labeled as such by professionals (rather than binary true vs. false labels).

[4] https://github.com/Tariq60/LIAR-PLUS

Statement:"Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations." **Speaker:** Florida Democratic Party **Context:** TV Ad **Label:** half-true **Extracted Justification:** A TV ad by the Florida Democratic Party says Scott "cut education to pay for even more tax breaks for big, powerful, well-connected corporations." However, the ad exaggerates when it focuses attention on tax breaks for "big, powerful, well-connected corporations." Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.

Table 5.12: Excerpt from the LIAR-PLUS dataset.

5.3.1 Dataset

The LIAR dataset introduced by Wang (2017) consists of 12,836 short statements taken from POLITIFACT and labeled by humans for truthfulness, subject, context/venue, speaker, state, party, and prior history. For truthfulness, the LIAR dataset has six labels: *pants-on-fire, false, mostly-false, half-true, mostly-true,* and *true*. These six label sets are relatively balanced in size. The statements were collected from a variety of broadcasting mediums, like TV interviews, speeches, tweets, and debates, and they cover a broad range of topics such as the economy, health care, taxes, and elections.

We extend the LIAR dataset to the LIAR-PLUS dataset by automatically extracting for each claim the justification that humans have provided in the fact-checking article associated with the claim. Most of the articles end with a summary that has the headline "our ruling" or "summing up". This summary usually has several justification sentences that are related to the statement. We extract all sentences in these summary sections, or the last five sentences in the fact-checking article when no summary exists. We filter out the sentence that has the verdict and related words. These extracted sentences can support or contradict the statement, which is expected to enhance the accuracy of the classification approaches. An excerpt from the LIAR-PLUS dataset is shown in Table 5.12.

5.3.2 Methods

Our main goal in this section is to show that modeling the human-provided justification — which can be seen as summary evidence — improves the assessment of a claim's degree of truthfulness when compared to modeling the claim (and metadata) alone, regardless of the machine learning models (feature-based vs. deep learning models). All our models use four different conditions: *basic claim/statement*[5] *representation* using just word representations (**S condition**), *enhanced claim/statement representation* that captures additional information shown to be useful such as hedging, sentiment strength and emotion (Rashkin et al., 2017a) as well as *metadata information* (**S$^+$M condition**), *basic claim/statement* and the associated *extracted justification* (**SJ condition**) and finally *enhanced claim/statement representation*, *metadata* and *justification* (**S$^+$MJ condition**).

Feature-based Machine Learning. We experiment with both Logistic Regression (LR) and Support Vector Machines (SVM) with a linear kernel. For the basic representation of the claim/statement (S condition), we experimented with unigram features, TF-IDF unigram features, and GLoVe word embeddings (Pennington et al., 2014). The best representation proved to be unigrams. For the enhanced statement representation (S$^+$) we modeled: sentiment strength using SentiStrength, which measures the negativity and the positivity of a statement on a scale of 1-to-5 (Thelwall et al., 2010); emotion using the NRC Emotion Lexicon (EmoLex), which associates each word with eight basic emotions (Mohammad and Turney, 2010), and the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001). In addition, we include metadata information such as the number of claims each speaker makes for every truth label (history) (Wang, 2017; Long et al., 2017). Finally, for representing the justification in the SJ and S$^+$MJ conditions, we just use unigram features.

Deep Learning Models. We use a Bi-Directional Long Short-term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997) architectures that has been shown to be successful for various related NLP tasks such as textual entailment and argument mining. For the S condition, we use just one

[5]referred to as statement henceforth in this section.

Cond.	Model	Binary		Six-way	
		valid	test	valid	test
S	LR	0.58	0.61	0.23	0.25
	SVM	0.56	0.59	0.25	0.23
	BiLSTM	0.59	0.60	0.26	0.23
SJ	LR	0.68	0.67	0.37	0.37
	SVM	0.65	0.66	0.34	0.34
	BiLSTM	0.70	0.68	0.34	0.31
	P-BiLSTM	0.69	0.67	0.36	0.35
S^+M	LR	0.61	0.61	0.26	0.25
	SVM	0.57	0.60	0.26	0.25
	BiLSTM	0.62	0.62	0.27	0.25
S^+MJ	LR	0.69	0.67	0.38	0.37
	SVM	0.66	0.66	0.35	0.35
	BiLSTM	0.71	0.68	0.34	0.32
	P-BiLSTM	0.70	0.70	0.37	0.36

Table 5.13: Classification results.

BiLSTM to model the statement. We use GLoVe pre-trained word embeddings (Pennington et al., 2014), a 100-dimensional embedding layer that is followed by a BiLSTM layer of size 32. The output of the BiLSTM layer is passed to a softmax layer. In the S^+M condition, a normalized count vector of those features (described above) is concatenated with the output of the BiLSTM layer to form a merged layer before the softmax. We use the categorical cross-entropy loss function and ADAM optimizer (Kingma and Ba, 2014) and train the model for 10 epochs. For the SJ and S^+MJ conditions, we experiment with two architectures: in the first one, we just concatenate the justification to the statement and pass it to a single BiLSTM, and in the second one we use a dual/parallel architecture where one BiLSTM reads the statement and another one reads the justification (architecture denoted as P-BiLSTM). The outputs of these BiLSTMs are concatenated and passed to a softmax layer. This latter architecture has been proven to be effective for tasks that model two inputs such as textual entailment (Conneau et al., 2017b) or sarcasm detection based on conversation context (Ghosh et al., 2017; Ghosh and Veale, 2017).

Class	class size	S		SJ		
		LR	BiLSTM	LR	BiLSTM	P-BiLSTM
pants-fire	116	0.18	0.19	0.37	0.34	0.37
false	263	0.28	0.34	0.33	0.3	0.33
mostly-false	237	0.21	0.13	0.35	0.31	0.32
half-true	248	0.22	0.28	0.39	0.31	0.37
mostly-true	251	0.23	0.33	0.40	0.39	0.39
true	169	0.22	0.18	0.37	0.42	0.39
total/avg	1284	0.23	0.26	0.37	0.34	0.36

Table 5.14: F1 score per class on validation set.

Class	class size	S		SJ		
		LR	BiLSTM	LR	BiLSTM	P-BiLSTM
pants-fire	92	0.12	0.11	0.38	0.33	0.39
false	250	0.31	0.31	0.35	0.32	0.35
mostly-false	214	0.25	0.15	0.35	0.27	0.33
half-true	267	0.24	0.26	0.41	0.27	0.34
mostly-true	249	0.23	0.30	0.35	0.35	0.33
true	211	0.25	0.16	0.37	0.36	0.41
total/avg	1283	0.25	0.23	0.37	0.31	0.35

Table 5.15: F1 score per class on test set.

5.3.3 Results and Error Analysis

Table 5.13 shows the results both for the binary and the six-way classification tasks under all 4 conditions (S, SJ, S^+M and S^+MJ) for our feature-based machine learning models (LR and SVM) and the deep learning models (BiLSTM and P-BiLSTM).

For the binary runs, we group *pants on fire, false and mostly false* as FALSE and *true, mostly true and half true* as TRUE. As a reference, Wang (2017) (best models (text and metadata) obtained 0.277 F1 on the validation set and 0.274 F1 on the test set in the six-way classification, showing relatively similar results with our equivalent S^+M condition.

It is clear from the results shown in Table 5.13 that including the justification (SJ and S^+MJ conditions) improves over the conditions that do not use the justification (S and S^+M, respectively) for all models, both in the binary and the six-way classification tasks. For example, for the six-way

classification, we see that the BiLSTM model for the SJ condition achieves 0.35 F1 compared to 0.23 F1 in the S condition. LR model has a similar behavior with 0.37 F1 for the SJ condition compared to 0.25 F1 in the S condition. For the S^+MJ conditions, the best model (LR) shows an F1 of 0.38 compared to 0.26 F1 in the S^+M condition (similar results for the deep learning). The dual/parallel BiLSTM architecture yields a small improvement over the single BiLSTM only in the six-way classification.

We also present the per-class results for the six-way classification for the S and SJ conditions. Table 5.14 shows the results on the validation set, while Table 5.15 on the test set. In the S condition, we see a larger degree of variation in performance among the classes, with the worst being the pants-on-fire for all models, and for the deep learning model also the *mostly-false and true* classes. In the SJ condition, we notice a more uniform performance on all classes for all the models. We notice the biggest improvement for the pants-on-fire class for all models, *half-true* for LR, and *mostly-false* and *true* for the deep learning models. When comparing the P-BiLSTM and BiLSTM, we noticed that the biggest improvement comes from the half-true class and the pants-on-fire class.

Error Analysis In order to further understand the cause of the errors made by the models, we analyze several examples by looking at the statement, the justification, and the predictions by the logistic regression model when using the S, S^+M, SJ, and S^+MJ conditions (Table 5.16). Logistic regression is selected since it performs best for the six-way classification task.

The first example in Table 5.16 is wrongly classified in the S condition but correctly classified in the S^+M, SJ, and S^+MJ conditions. The justification text has a sentence saying "Statutory income tax rates in the U.S. fall around the end of the upper quarter of nations.", which contradicts the statement and thus is classified correctly when modeling the justification.

The second and the third examples in Table 5.16 are correctly predicted only when the justification was modeled (SJ and S^+MJ conditions). For statement 2, the justification text has a sentence "However, the ad exaggerates..." indicates that the statement has some false and some true information. Therefore, the model predicts the correct label "half-true" when modeling the

ID	Statement	Justification	label	S	S$^+$M	SJ	S$^+$MJ
1	We have the highest tax rate anywhere in the world.	Trump, while lamenting the condition of the middle class, said the U.S. has "the highest tax rate anywhere in the world." All sets of data we examined for individual and family taxes prove him wrong. Statutory income tax rates in the U.S. fall around the end of the upper quarter of nations. More exhaustive measures - which compute overall tax burden per person and as a percentage of GDP - show the U.S. either is in the middle of the pack or on the lighter end of taxation compared with other advanced industrialized nations.	false	X	✓	✓	✓
2	“Says Rick Scott cut education to pay for even more tax breaks for big, powerful, well-connected corporations.”	A TV ad by the Florida Democratic Party says Scott "cut education to pay for even more tax breaks for big, powerful, well-connected corporations." However, the ad exaggerates when it focuses attention on tax breaks for "big, powerful, well-connected corporations." Some such companies benefited, but so did many other types of businesses. And the question of whether the tax cuts and the education cuts had any causal relationship is murkier than the ad lets on.	half-true	X	X	✓	✓
3	Says Donald Trump has given more money to Democratic candidates than Republican candidates.	but public records show that the real estate tycoon has actually contributed around $350,000 more to Republicans at the state and federal level than Democrats. That, however, is a recent development. Ferguson’s statement contains an element of truth but ignores critical facts.	mostly-false	X	X	✓	✓
4	Says out-of-state abortion clinics have marketed their services to minors in states with parental consent laws.	As Cousins’ clinic in New York told Yellow Page users in Pennsylvania, "No state consents." This is information the clinics wanted patients or potential patients to have, and paid money to help them have it. Whether it was to help persuade them to come in or not, it provided pertinent facts that could help them in their decision-making. It fit the definition of marketing.	true	X	X	X	✓
5	Obamacare provision will allow forced home inspections by government agents.	But the program they pointed to provides grants for voluntary help to at-risk families from trained staff like nurses and social workers. What bloggers describe would be an egregious abuse of the law — not what’s allowed by it.	pants-fire	X	X	X	✓
6	In the month of January, Canada created more new jobs than we did.	In November 2010, the U.S. economy created 93,000 jobs, compared to 15,200 for Canada. And in December 2010, the U.S. created 121,000 jobs, compared to 22,000 for Canada. "But on a per capita basis, in recent months U.S. job creation exceeded Canada’s only in October." January happened to be a month when U.S. job creation was especially low and Canadian job creation was especially high, but it is the most recent month and it reflects the general pattern when you account for population.	true	X	X	X	X
7	There has been $5 trillion in debt added over the last four years.	number is either slightly high or a little low, depending on the type of measurement used, and that’s actually for a period short of a full four years. His implication that Obama and the Democrats are to blame has some merit, but it ignores the role Republicans have had.	mostly-true	X	X	X	X

Table 5.16: Error analysis of six-way classification (logistic regression).

justification text. Also, the justification for statement 3 was simple enough for the model to predict the gold label “mostly-false”. It has a phrase like “more to Republicans”, while the statement had “more to Democratic candidates”, which indicates falsehood in the statement as well as discourse markers indicating concessive moves (“but” and “however”).

Sometimes justification features alone are not enough to get the correct prediction without using

the *enhanced statement* and the metadata features. The justification for statement 4 in Table 5.16 is complex and no direct connection can be made to the statement. Therefore, the model fails when using SJ and S^+M conditions and only succeeds when using all features (i.e., S^+MJ condition). In addition, consider the fifth statement in Table 5.16 about Obamacare: it seems that metadata features, which have the history of the speaker, might have helped in predicting its factuality to be *pants on fire*, while it is wrongly classified when modeling only the statement and the justification.

For around half of the instances in the validation set, all models had wrong predictions. This is not surprising since the best model had an average F1 score of less than 0.40. The last two examples in Table 5.16 are instances where the model makes mistakes under all four conditions. The claim and the justification refer to temporal information, which is harder to model by the rather simple and shallow approaches we used. Incorporating temporal and numeric information when modeling the claim and the justification would be essential for capturing the correct context of a given statement. Another source of errors for justification-based conditions is the noise in the extraction of the justification, particularly when the "our ruling" and "summing up" headers are not included and we resort to extracting the last five sentences from the fact-checking articles. Improving the extraction methods will be helpful in improving the justification-based classification results.

5.4 Conclusion

In this Chapter, we presented models and attacks for end-to-end fact-checking and discussed the role of justification and fallacies for fine-grained claim verification.

We presented in Section 5.1 one of the first end-to-end systems for fact extraction and verification that was ranked sixth (out of 24) in the first FEVER shared task in 2018. The system consists of three components that include: i) document retrieval using a combined retrieval method over Google custom search, named entity recognition, and dependency parsing; ii) evidence sentence selection using TD-IDF and ELMo embeddings. iii) claim verification using the InferSent model for textual entailment.

We then discussed in Section 5.2 advancement in models for automatic fact-checking and showed the vulnerabilities of those models under three main categories of adversarial attacks that include: i) multi-hop propositions, ii) temporal reasoning, and iii) lexical variations.

In Subsection 5.3, we studied fine-grained truth labels of six truth levels that are more common in naturally occurring texts that include misleading statements and ones taken out of context thus having characteristics of misinformation that increase the complexity of their verification. We showed that using evidence through human-provided justifications is crucial for this task regardless of the machine learning model used whether it is a feature-based linear model or a neural network. This empirically shows the importance of conducting evidence-based verification of naturally occurring claims.

Chapter 6

Conclusions

In this thesis, we studied the role of argument structure and argument quality in improving tasks related to fact-checking and (mis-/dis-)information detection. We covered a wide range of tasks related to misinformation detection and fact-checking, such as distinguishing factual statements from opinions, assessing the checkworthiness of information in news articles, studying the connection between fallacies and misinformation and proposing a unified model for fallacy recognition and developing claim verification approaches given automatically retrieved (or provided) evidence under truth barometers with different levels.

We were able to predict whether a news article is a news story or an opinion pieces through argumentation features based on predicted types of argument components in the articles. We also showed the role of argument structure for checkworthiness by using gold annotations of the argument structure to inform the selection of a more useful context for checkworthiness prediction.

We investigated fallacies as indicators of misinformation and developed models for fallacy recognition in single- and multi-dataset settings. We showed the resilience of multitask instruction-based prompting for fallacy recognition across four fallacy schemes and five fallacy datasets that cover multiple domains and genres. This approach still suffers from recognizing fallacies that might require external knowledge such as diversion fallacies, but it was able to provide useful explanations for the checkworthiness of statements on climate change and Covid-19 in more than 64% of the examples.

We then presented our fact extraction and verification models that cover a number of steps, including evidence document retrieval, evidence sentence selection and claim verification. We also studied the robustness of fact-checking models under adversarial attacks. Finally, we showed that

evidence is essential for fine-grained verification of naturally occurring claims by modeling the human-provided justifications.

6.1 Contributions

We restate our main contributions below.

- We approached fact-checking with a holistic view by developing models for checkworthiness (what to fact-check), fallacy recognition (why to fact-check), and veracity prediction (how to fact-check), in addition to analyzing the connections between these tasks.
- We utilized features from the argument structure in two downstream tasks:
  - separating facts from opinions in news articles
  - predicting the checkworthiness of statements in news articles
- We presented models for fallacy recognition trained under different settings and studied their role in explaining checkworthiness
  - We presented a new scheme for fallacy in fact-checked content in collaboration with (Musi et al., 2022).
  - We introduced a unified model for fallacy recognition using multitask instruction-based prompting
  - We use fallacies as an indicator of checkworthiness in climate change and Covid-19
- We presented models for end-to-end fact-checking using different truth barometers and evidence scenarios and studied their performance under adversarial attacks as follows:
  - We presented one of the first end-to-end models for fact extraction and verification that include finding relevant evidence to a given claim and assessing the veracity of the claim compared to the retrieved evidence.

  - We presented a number of adversarial attacks that show the ability of fact-checking models to handle multiple propositions, temporal reasoning, and lexical variations.
  - We showed the importance of evidence for fine-grained claim verification using human-provided justifications.

- We released a number of new datasets:
  - A multi-layer annotated corpus for checkworthiness and argumentative discourse structures for climate change news articles (Alhindi et al., 2021)
  - A fallacy corpus of climate change and Covid-19 news articles and social media posts (Alhindi et al., 2022).
  - The LIAR-PLUS dataset (Alhindi et al., 2018) of fact-checked claims with justification.

6.2 Limitations and Future Work

We discuss below the limitations of our work and potential future work in mining argument structure and extracting argumentation features, fallacy recognition application to fine-grained claim verification.

Argument Structure and Argumentation Features Argumentation features are good for document-level tasks with over-prediction of argumentative components due to the nature and distribution of the training data. To improve the prediction of argument components and relations, we could use an argumentation model on news articles similar to our multitask token-based argument segmentation and argument component type prediction model on essays (Alhindi and Ghosh, 2021). Other models could be used as well that cover end-to-end argumentation (Eger et al., 2017) or use discourse to inform argumentation mining (Saha et al., 2022). In addition, to limit the over-prediction of argumentative components, we can sample non-argumentative sentences in news articles and train on a more balanced dataset of argumentative and non-argumentative sentences. Also, adding a

mix of topics to the training is expected to increase the accuracy of the predictions and enable using them on more fine-grained downstream tasks and not only document-level classification. Additionally, further research in argument modeling for long-form text is needed to better chapter relations between components that are further away in the text. This could allow methods such as ours for checkworthiness prediction using argumentation context to be applied to datasets with no annotations of argument structures.

Fallacies and Shades of Truth We have shown in Section 5.3 a six-level claim veracity rating scheme by PolitiFact that does not follow a clear-cut distinction between *Supported/True* and *Refuted/False* as the one introduced in the (synthesized) FEVER dataset. In fact, most fact-checking organizations use multi-level or multi-facet schemes that capture nuanced relations between the claim and the evidence. For example, SNOPES[1] has labels such as *outdated, unproved*, and *misattributed*, and SCIENCEFEEDBACK[2] has ones like *lacks-context*, and *inaccurate*. Moreover, FULLFACT[3] refrains from giving any kind of verdict and only provides a summary sentence (or two) that explains what is wrong with a particular claim. This clearly indicates the complexity of capturing the relation between a claim and evidence in naturally occurring claims and shows the need to better understand the difference between these more fine-grained labels. In future work, we could utilize our fallacy recognition model for the claim veracity prediction task and study the relation between labels like *half-true* and fallacy types. This work could be done on the LIAR/LIAR-PLUS dataset that has the six-degree truth labels. The LIAR dataset is also from the same genre of our Covid-19 fallacy data that has claims from POLITIFACT as well. Although the two datasets are on different topics, we hypothesize that argumentative fallacies could provide informative inputs for nuanced prediction of claim veracity.

[1]https://www.snopes.com/fact-check-ratings/
[2]https://sciencefeedback.co/claim-reviews-framework/
[3]https://fullfact.org/about/frequently-asked-questions/ratings

Bibliography

Sean Adl-Tabatabai. 2016. Tens of thousands of scientists declare climate change a hoax - your news wire. https://web.archive.org/web/20171015212815/http://yournewswire.com/tens-of-thousands-of-scientists-declare-climate-change-a-hoax/. (Accessed on 10/14/2022).

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357.

Amal Alabdulkarim and Tariq Alhindi. 2019. Spider-Jerusalem at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 985–989, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Tariq Alhindi and Debanjan Ghosh. 2021. Sharks are not the threat humans are: Argument component segmentation in school student essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, (Online). Association for Computational Linguistics.

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, November. Association for Computational Linguistics.

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 98–102, Hong Kong, China, November. Association for Computational Linguistics.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: The role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Tariq Alhindi, Brennan McManus, and Smaranda Muresan. 2021. What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure. In *The 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (submitted)*. Association for Computational Linguistics.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October-November. Association for Computational Linguistics.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. 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, November. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. 2017. Detecting Good Arguments in a Non-Topic-Specific Way: An Oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249, Vancouver, Canada. Association for Computational Linguistics.

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 109(4):719–760.

Bettina Berendt, Peter Burger, Rafael Hautekiet, Jan Jagers, Alexander Pleijter, and Peter Van Aelst. 2020. FactRank: Developing automated claim detection for dutch-language fact-checkers.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jonathan S Blake et al. 2019. *News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms*. Rand Corporation.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation*. Ph.D. thesis, University of Oxford.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131, Brussels, Belgium, November. Association for Computational Linguistics.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. Bllip 1987-89 wsj corpus release 1. *Linguistic Data Consortium, Philadelphia*, 36.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017b. Reading wikipedia to answer opendomain questions. pages 1870–1879. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the*

Fourteenth Workshop on Semantic Evaluation, pages 1462–1468, Barcelona (online), December. International Committee for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification*, FEVER '18, pages 40–49, Brussels, Belgium.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017b. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China, November. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5636–5646, Hong Kong, China, November. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December. International Committee for Computational Linguistics.

T Edward Damer. 2012. *Attacking faulty reasoning*. Cengage Learning.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online, August. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the

persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! lab: automatic identification and verification of claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 301–321. Springer.

George Englebretsen. 1973. Fallacies. by cl hamblin. london: Methuen. 1970. pp. 326. 5.65 (paperback). *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, 12(1):151–154.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Matt Gardner, Joel Grus, Oyvind Tafjord Mark Neumann, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform.

Tedros Adhanom Ghebreyesus. 2020. Munich security conference. https://www.who.int/director-general/speeches/detail/munich-security-conference. (Accessed on 10/14/2022).

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Debanjan Ghosh, R. Alexander Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization, 7. Main Track.

Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. Technical report, Reuters Institute, University of Oxford.

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

Pankaj Gupta, Khushbu Saxena, Usama Yaseen, Thomas Runkler, and Hinrich Schütze. 2019. Neural architectures for fine-grained propaganda detection in news. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 92–97, Hong Kong, China, November. Association for Computational Linguistics.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. UBY: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana, June. Association for Computational Linguistics.

Charles L Hamblin. 1970. Fallacies.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri,

Christian M Meyer, and Iryna Gurevych. 2018a. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 1859–1874, Santa Fe, New Mexico, USA.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018b. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium, November. Association for Computational Linguistics.

Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *CLEF*.

Hans V Hansen. 1996. Aristotle, whately, and the taxonomy of fallacies. In *International Conference on Formal and Applied Practical Reasoning*, pages 318–330. Springer.

Hans Vilhelm Hansen. 2002. The straw thing of fallacy theory: the standard definition of'fallacy'. *Argumentation*, 16(2):133–155.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States, July. Association for Computational Linguistics.

Laurie Beth Harris. 2017. Helping readers tell the difference between news and opinion: 7 good questions with duke reporters' lab's rebecca iannucci, Aug.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Christopher Hidey and Mona Diab. 2018. Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction*

and VERification (FEVER), pages 150–155, Brussels, Belgium, November. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online, July. Association for Computational Linguistics.

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

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4):1621.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana, June. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya

Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2018. Joint multitask learning for community question answering using task-specific embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4196–4207.

Garth S Jowett and Victoria O'Donnell. 2012. What is propaganda, and how does it differ from persuasion. *Propaganda & persuasion*, pages 1–48.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online), December. International Committee for Computational Linguistics.

Yavuz Selim Kartal, Busra Guvenen, and Mucahid Kutlu. 2020. Too many claims to fact-check: Prioritizing political claims based on check-worthiness. *arXiv preprint arXiv:2004.08166*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Youngwoo Kim and James Allan. 2019. FEVER breaker's run of team NbAuzDrLqg. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 99–104, Hong Kong, China, November. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey.

In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Katarina R Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistic*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium, November. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*.

Kevin Meng, Damian Jimenez, Fatma Arslan, Jacob Daniel Devasier, Daniel Obembe, and Chengkai Li. 2020. Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims. *arXiv preprint arXiv:2002.07725*.

Clyde R. Miller. 1939a. *The Techniques of Propaganda. From "How to Detect and Analyze Propaganda,"*. an address given at Town Hall. The Center for learning.

Clyde R. Miller. 1939b. *The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning*.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 767–776, New Orleans, Louisiana, USA.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-

2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online), December. International Committee for Computational Linguistics.

Elena Musi and Chris Reed. 2022. From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, page 09579265221076609.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland, June. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387. Springer.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021b. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *European Conference on Information Retrieval*, pages 639–649. Springer.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Mucahid Kutlu,

Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, and Alex Nikolov. 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets.

Callistus Ireneous Nakpih and Simone Santini. 2020. Automated discovery of logical fallacies in legal argumentation. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 11.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019b. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China, November. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy, July. Association for Computational Linguistics.

Ankur Parikh, Dipanjan Das Oscar Tackstr ¨ om, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. pages 2249–2255. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal, June. to appear*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018a. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018b. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

D. Pomerleau and D. Rao. 2017a. Fake news challenge. `http://www.fakenewschallenge.org/`. (Accessed on 12/06/2019).

Dean Pomerleau and Delip Rao. 2017b. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL (1)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017a. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017b. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2022. Newsclaims: A new benchmark for claim detection from news with background knowledge. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy, July. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450.

Victoria L Rubin, Yimin Chen, and Nadia K Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sougata Saha, Souvik Das, and Rohini K Srihari. 2022. Edu-ap: Elementary discourse unit based argument parser. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 183–192.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake

news detection with pre-trained transformer models. In *Proceedings of the ICPR International Workshops and Challenges*, pages 627–641.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020. Overview of CheckThat! 2020 english: Automatic identification and verification of claims in social media. In *CLEF (Working Notes)*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online, June. Association for Computational Linguistics.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain, April. Association for Computational Linguistics.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. In *Proceedings of the 2020 Truth and Trust Online Conference (TTO 2020)*, page 32. Hacks Hackers.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhance-

ment for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China, November. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China, November. Association for Computational Linguistics.

The-Media-Insight-Project. 2018. Americans and the news media: What they do — and don't — understand about each other.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558.

James Thorne and Andreas Vlachos. 2017. An extensible framework for verification of numerical claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain, April. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018c. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium, November. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China, November. Association for Computational Linguistics.

Christopher W Tindale. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Santosh Tokala, G Vishal, Avirup Saha, and Niloy Ganguly. 2019. Attentivechecker: A bi-directional attention flow mechanism for fact verification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2218–2222.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming.

Frans H Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301.

Frans H Van Eemeren, A Francisca Sn Henkemans, and Rob Grootendorst. 2002. *Argumentation: Analysis, evaluation, presentation*. Routledge.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, Maryland, USA.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International*

Conference on Computational Linguistics: Technical Papers, pages 1680–1691, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, April. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, November. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. pages 525–533. WWW '18 Companion Proceedings of the The Web Conference 2018.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Clair Wardle. 2016. 6 types of misinformation circulated this election season - columbia journalism review. https://www.cjr.org/tow$_center/6_types_election_fake_news.php$. (Accessed on 10/14/2022).

Clair Wardle. 2017. Fake news. it's complicated. https://firstdraftnews.org/articles/fake-news-complicated/. (Accessed on 10/14/2022).

Clair Wardle. 2020. Understanding information disorder - first draft. https://firstdraftnews.org/long-form-article/understanding-information-disorder/. (Accessed on 10/14/2022).

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Fact check-worthiness detection as positive unlabelled learning. *arXiv preprint arXiv:2003.02736*.

Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-worthiness detection for improved scientific document understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1796–1807, Online, August. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium, October-November. Association for Computational Linguistics.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium, November. Association for Computational Linguistics.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China, November. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430.

Haoran Zhang and Diane Litman. 2020. Automated topical component extraction using neural network attention scores from source-based essay scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8569–8584, Online, July. Association for Computational Linguistics.

Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. 2019a. From stances' imbalance to their hierarchical representation and detection. In *Proceedings of the World Wide Web Conference*, WWW '19, pages 2323–2332, Lyon, France.

Yi Zhang, Zachary Ives, and Dan Roth. 2019b. Evidence-based trustworthiness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 413–423.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Appendix A: Fallacies in Pragma-Dialectical Perspective

Van Eemeren and Grootendorst (1987) consider fallacies as violations of the ten rules of critical discussion. Below we list the ten rules and their violations that map to fallacious moves.

Rule I *Parties must not prevent each other from advancing or casting doubt on standpoints*. Rule I applies to the confrontation stage of a critical discussion, and can be violated by both the protagonist and the antagonist. Possible violations and the corresponding fallacies are: banning standpoints, declaring standpoints sacrosanct, putting pressure on the opponent (*ad baculum* and *ad misericordiam*), or performing a personal attack on the opponent (*ad hominem*)

Rule II *Whoever advances a standpoint is obliged to defend it if asked to do so*. Rule II applies to the opening stage, and can be violated by the protagonist: *evading the burden of proof* by presenting the standpoint as self-evident, giving a personal guarantee of the rightness of the standpoint, or immunizing the standpoint against criticism, or by *shifting the burden of proof* by demanding the antagonist shows that the standpoint is wrong.

Rule III *An attack on a standpoint must relate to the standpoint that has really been advanced by the protagonist*. Rule III applies to all stages of a critical discussion, and can be violated by the antagonist: Imputing a fictitious standpoint to someone, or distorting someone's standpoint by *oversimplification* or *exaggeration*.

Rule IV *A standpoint may be defended only by advancing argumentation relating to that standpoint*. Rule IV applies to the argumentation stage, and can be violated by the protagonist: irrelevant argumentation (*ignoratio elenchi*), using pathos by playing on the emotions or prejudices of the audience (*argumentum ad populum*) and using ethos by parading one's own qualities

(*argumentum ad verecundiam*).

Rule V *A person can be held to the premises he leaves implicit*. Rule V applies to the argumentation stage, and can be violated by both the protagonist and the antagonist: reconstructing an unexpressed premise beyond what the protagonist can be held to (antagonist), denying a commitment to a correctly reconstructed unexpressed premise (protagonist).

Rule VI *A standpoint must be regarded as conclusively defended if the defence takes place by means of arguments belonging to the common starting point.* Rule VI applies to the argumentation stage and can be violated both by the protagonist and the antagonist: wrapping up a proposition in a presupposition (protagonist), hiding away a proposition in an unexpressed premise (protagonist), advancing an argument that amounts to the same thing as the standpoint (protagonist), or casting doubt on a starting point (antagonist).

Rule VII *A standpoint must be regarded as conclusively defended if the defence takes place by means of arguments in which a commonly accepted scheme of argumentation is correctly applied*. Rule VII applies to the argumentation stage, and can be violated by the protagonist: Applying an unsuitable scheme of argumentation such as the *appeal to irrelevant authority* or the *bandwagon* fallacy, Inappropriately applying a scheme of argumentation which includes fallacies such as *hasty generalization*, *false analogy*, *post hoc*, and *slippery slope*.

Rule VIII *The arguments used in a discursive text must be valid or capable of being validated by the explicitization of one or more unexpressed premises*. Rule VIII applies to the argumentation stage, and can be violated by the protagonist in various ways: (a) Confusion of necessary and sufficient conditions, (b) Confusion of properties of parts and wholes which includes *fallacy of division* and *fallacy of composition*.

Rule IX *A failed defence must result in the protagonist withdrawing his standpoint and a successful defence must result in the antagonist withdrawing his doubt about the standpoint.* Rule

IX applies to the concluding stage and can be violated by both the protagonist and the antagonist: concluding that a standpoint is true because it has been successfully defended against the opposition of the antagonist (by the protagonist), concluding that a standpoint is true because the opposite has not been successfully defended (by the antagonist), often combined with an erroneous opposition showcasing the fallacy of *false dilemma*.

Rule X *Formulations must be neither puzzlingly vague nor confusingly ambiguous and must be interpreted as accurately as possible*. Rule X applies to all the stages of a critical discussion, and can be violated by both the protagonist and the antagonist. The main types of unclearness and ambiguity are: structural unclearness (textual level), implicit illocutionary force of a speech act (sentence level), indefinite reference of a speech act (sentence level), unfamiliar predication of speech act (sentence level), vague predication of a speech act (sentence level), semantic ambiguity, and syntactic ambiguity.

More details about the stages and rules of critical discussion and examples for each violation can be found in Van Eemeren and Grootendorst (1987).