Classifying Stance in News Articles: Use of Attribution Relations and Source Expertise

Thesis

School of Computing and Communications

Classifying Stance in News Articles: Use of Attribution Relations and Source Expertise

Supervisors:
Dr. Alistair Willis
Dr. Paul Piwek

Author:
Nitu Bharati

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Artificial Intelligence and Natural Language Processing
2023

Abstract

The overarching aim behind this research is to automatically detect the stance of the body of a news article relative to the article's headline. The news headline may not always reflect what is in the news body. The stance of a news body to its headline can be agree, disagree, discuss or unrelated (Pomerleau & Rao 2017). Central to this work is the use of a specific discourse relation, the attribution relation (AR), for detecting the stance of a news article body relative to its headline. An attribution relation is a span of text which links a source to content through a cue. For example, consider *The boy said it was a spider.* Here, *the boy* is the source, *said* is the cue and *it was a spider* is the content. This thesis also examines how the expertise of sources affects stance detection. The main research question of this work is "Can attribution relations and source expertise be useful in detecting the stance of a news article's body towards its headline?".

To address this research question, I developed a new attribution detection model that can tag components of attribution relations in news texts. I developed a new stance detection model which uses these tags as input, rather than working on the whole article as a single piece of text, with performance comparable to state-of-the-art. Furthermore, once we add the source expertise information to our stance detection model, this has a positive effect on the F-score for stance detection (increase by 14%).

The work is novel in a number of further specific ways. Firstly, it is the first time a single-step deep learning approach has been applied to AR detection and been released as open source code. Second, this is the first time that attribution relations from a news article body have been used as input for a stance detection model instead of the full text of the news article body. As part of this research I created an extension to the Fake news challenge corpus (Pomerleau & Rao 2017) with addition of source expertise data. Finally, I separately confirmed, through an empirical study, that source expertise is positively correlated with the credibility that readers assign to claims from a source.

Acknowledgement

Firstly I express my sincere gratitude to my supervisors Alistair Willis and Paul Piwek for their constant support and encouragement to pursue my research successfully.

I express my sincere appreciation and graceful thanks to the Open University for providing me with this opportunity with necessary funding and infrastructure for my research and very helpful PG forums and NLP group meetings.

I sincerely thank the Human Research Ethics Committee in the Open University for assuring the ethical integrity of my research related to an empirical study and corpus building. I would like to thank Alvaro Faria from the School of Mathematics and Statistics for an expert statistical advisory service. I express my sincere thanks to my friends and colleague Jacopo and German for their participation and feedbacks on a pilot study carried out during my research.

I would like to thank Bobo, Matt, Sonia, Binita, Suresh, Jana, Yadu and Telma who made this journey pleasant. It is never enough thanking my partner Suraj Jung Pandey who was constant source of happiness, even during tough times of pandemic.

Thanking is never enough for your unconditional love and support through out the journey: Dad, mom, sisters, brother, nephew and nieces I love you all a lot.

Dedicated to my parents:

Balabhadra Bharati and Chandra Bharati

Contents

List of Figures

List of Tables

Glossary

attribution relation A discourse relation which connects an attributional object (content) to the owner of the attribution (source) through an attributional anchor (cue)
For example: *The boy said it was a spider.*
Here *the boy* is source
said is cue
it was a spider is content. 8, 23, 27–29, 58, 124

content The attributional object of an attribution relation. The content consists of information bearing text (which the attribution relation attributes to its source) (Pareti 2012) . xiv, 30

cue An anchor of speech act or cognitive predicate that attributes the content to its source (Pareti 2012). xiv, 30

disinformation Intentional sharing of false information to harm others. 5, 6, 21

malinformation True information shared to cause harm, often by moving information designed to stay private into the public sphere . 5

misinformation False information shared with no intention to harm others. 5

proposition Meaning of a sentence occurring from a speech act or cognition (McGrath & Frank 2018). xv

propositional attitude Cognitive relation an entity bears to a proposition (Nelson 2019) . 29

source A communicative agent or an artifact (Newell, Margolin & Ruths 2018). . xiv, 30

stance detection The automated task of identifying the perspective of a text (news body) relative to another text (news headline), specifically in terms of the main claim conveyed by the headline.. 1, 14, 15, 28, 57, 124

veracity Truthfulness or being true. 15

veracity assessment Assessing the truthfulness of a specific claim. 15, 17

Chapter 1

Introduction

In everyday language the *stance* of a text, informational item (diagram) or even person is a relation to a claim, issue or topic. Pomerleau & Rao (2017) defined stance detection in news articles as "the task of estimating the relative perspective of two pieces of text relative to a topic, claim or issue". In this work I interpret stance detection in news articles as the automated task of identifying the perspective of a text (news body) relative to another text (news headline), specifically in terms of the main claim conveyed by the headline. In a news article, the news body can be agreeing, disagreeing, discussing or be irrelevant to the headline. Let us consider the following example,

> **Example 1.1.**
> **Headline**: Tropical spider burrowed under man's skin through appendix scar and lived there for THREE DAYS
>
> **News body**: Prepare to meet ... mite man. Doctors removed a matchhead-sized insect, believed to be a spider, from under Dylan Thomas's skin earlier this week and have sent the creature away for testing to determine what it is. It had been there for three days and burrowed up to his chest, leaving a trail of red blisters. The 21-year-old was on his first trip to Bali. He told

> News Corp yesterday that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin. There's just one problem. Spiders, according to Perth arachnid expert Dr Volker Framenau, don't burrow in skin. "They don't have the tools, the armature, to do this sort of stuff," Dr Framenau said. "I find it highly unlikely, almost impossible, that it was a spider.' More likely, Dr Framenau said, was some kind of burrowing mite. "That's a professional skin-digger," he said. "There's a lot of nasty stuff out there." The results of the tests on the creature should come back next week. Mr Thomas has been contacted for comment.

In the news body of Example 1.1, there is a boy who is claiming that doctors pulled out a spider from his body. The boy's claim in the news body is agreeing with the claim in the headline. In contrast, there is an arachnid expert in the news body who is claiming that it cannot be a spider because spiders don't have the tools to burrow under the skin. The expert's evidence on the spider being incapable of burrowing under human skin means the whole news body disagreeing with the claim in the headline. Hence, here the stance of the news body to its headline is disagree. I am interested to find the stance between the news body and the headline in my research. In addition, I will also do a fine-grained analysis of news body contents and their role in the stance detection.

In news reporting, a headline is generally a short text summarising an associated news article. Typically a reader expects the headline to be an accurate reflection of the article (Dor 2003). It is unusual to find any deviation in a news body from what is claimed in the headline. However, such deviations can occur due to sensational headlines used for news reporting (Molek-Kozakowska 2013), to increase clicks (Abhijnan Chakraborty 2016) or deliberately done to spread misinformation (Silverman 2015). Gabielkov et al. (2016) found that more than half of people share news articles on social media without clicking the link to the actual article. People often share news articles only by reading their headlines if the headline contents agree with an existing belief system of people, known as confirmation bias (Knobloch-Westerwick & Kleinman 2012), which may lead to the spread of misinformation. A news headline-body dissonance capturing technique like stance detection is needed to flag such news articles. However, there are many challenges to detect stance in news articles such as lack of a balanced dataset with equal representative samples and techniques to capture subtle deviations in news articles. Apart from its need in misinformation domain, the stance detection is used to understand argumentative structure in persuasive essays (Stab & Gurevych 2014). Stab & Gurevych (2014) identified whether the relation is supporting or not supporting between argumentative components like major claims, claims and premises. Somasundaran & Wiebe (2010) mentioned that the stance detection is useful to analyse ideological online debates. Somasundaran & Wiebe (2010) detected the overall position taken by a person respective to a given topic based on his/her arguing expressions of opinion and sentiment.

Further on stance detection applications, it has potential to be applicable in various other fields. Stance detection is used to analyse political topics for instance, analysing public reactions towards Brexit using Twitter data (Grčar et al. 2017). Furthermore, stance detection is used to analyse the diachronic evolution of people's views over time (Alkhalifa et al. 2021). In the analysis, the stance toward a specific target is classified at the user level by aggregating data over time, considering different time-window sizes and using temporally adapted word embeddings to re-train the classifier on the unchanged training data (Alkhalifa et al. 2021). A further application of the stance detection includes identifying political ideology of a person using an image content (Xi et al. 2020). Xi et al. (2020) showed that features in an image related to patriotism (like the country's flag), military (such as military band), economic inequalities and minority groups are useful to classify a person's ideology in the image as Republic or Democratic.

There are many potential applications of my research that is also further discussed in Section 8.3. Attribution relations detected in a news body can be used to get a glance insight of the news article. A reader can have a quick look into all important bit of information from a news article in the form of attribution relations, avoiding all background details and writer's opinions in the news article. A work by Anastasiou & De Liddo (2021) showed that automated reports like arguments and summaries are useful to improve sensemaking and perceived quality of a long content like debate. Furthermore, such reports are useful to provide a quick insight of a long content because such reports are short and manageable than the actual long content. So, attribution relations could be useful to get a good sense about the news article. Furthermore on the potential benefit of my research, my stance detection model could be useful for selecting a relevant headline for a given news text. Dor (2003) highlighted that the most appropriate headline for a news item is the one which optimizes the relevance of the story for the readers of the newspaper. For this, we can use an off-the-shelf tool to generate several headlines for a given text. My stance detection model can be useful to find with which headline the given text is agreeing or disagreeing. Amongst headlines with which a news body is agreeing, we can select the most relevant headline for the news.

Stance detection is relevant but, different from the task of incongruence/dissonance detection. Chesney et al. (2017) defined incongruent headlines as the ones that do not accurately represent the article information with which they appear. With this definition, I believe incongruent headlines has similarity to the headline-bodies that are in a disagreement or are irrelevant. However, Chesney et al. (2017) argued that the disagree class in the Fake news challenge corpus (Pomerleau & Rao 2017) represents a direct and strong contradiction between the news headline and body pairs. In contrast, Chesney et al. (2017) argued that incongruence is subtle exaggeration or misrepresentation of information that might not necessarily represent an opposing view. Another work by Park et al. (2020) mentioned that stance detection and headline incongruence problem are technically related as they both consider textual relationship between a headline and the associated news body. However, congruent and incongruent headlines cannot be mapped directly to the respective stance classes related (agree, disagree and discuss) and unrelated in the Fake News Challenge corpus. Park et al. (2020) argued that stance detection data can not be used for the headline incongruence problem because headline-body pairs with a related stance can be incongruent. In contrast to these works, Kumar et al. (2022) used respective related and unrelated data of stance detection task (Pomerleau & Rao 2017) as congruent and incongruent class data in the headline incongruence problem. Kumar et al. (2022) presented stance detection as the incongruence news detection problem. Although Kumar et al. (2022) claimed that their approach solves the headline incongruence problem, it is actually solving a different problem.

Finding the stance between the headline and body is an important aspect of sharing news in social media. The spread of misinformation can happen by sharing news headlines that are not in a complete agreement with their associated news bodies (Silverman 2015). As many people share news on social media without reading the whole news article (Gabielkov et al. 2016), such news articles may have news bodies not reflected by their headlines. The stance label of a news article might help readers to make an informed judgement about its sharing. Therefore, in my work, I classify the stance of the news body to its headline. The stance helps readers to spot if there exists any disagreement between the news headline and the body.

Misinformation not only affects an individual but can have an adverse effect in the population as a whole. In 2014, the World Economic Forum asked members of its Network of Global Agenda Councils to identify and prioritize the issues that could highly impact the world in future. The Network of Global Agenda Councils listed *The rapid spread of misinformation online* in the first ten. *The rapid spread of misinformation online* was named along with global critical issues such as inaction on climate change (Forum 2014).

There are various terms relating to misinformation. Wardle & Derakhshan (2017) have come up with an influential definition that distinguishes misinformation, disinformation and malinformation. The definition is based on

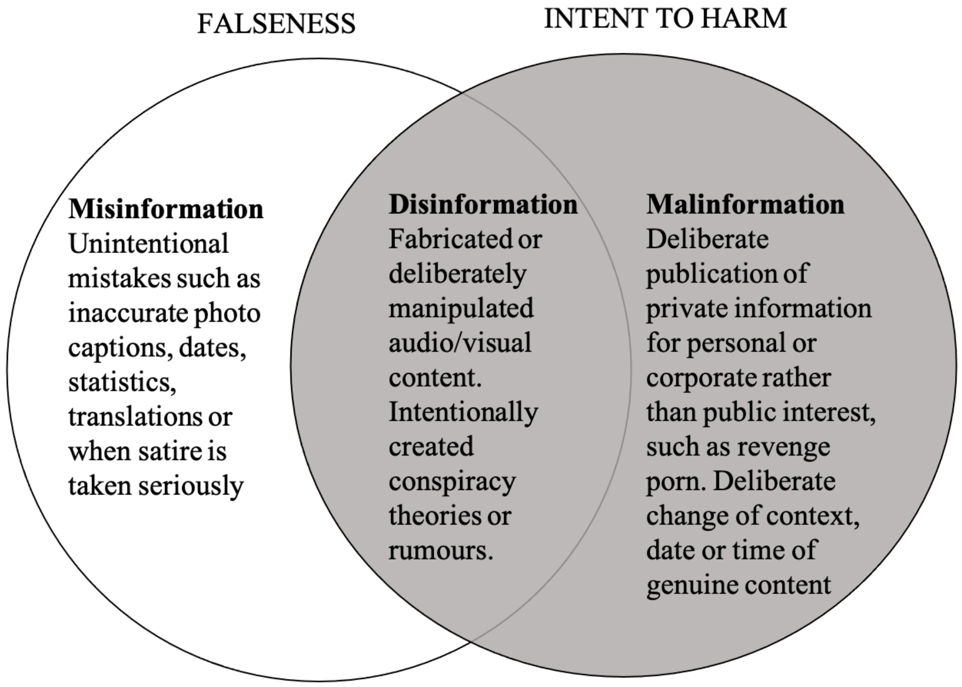

FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as inaccurate photo captions, dates, statistics, translations or when satire is taken seriously
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content

Figure 1.1: Information Disorder Framework, adapted from Wardle & Derakhshan (2017)

the two dimensions *falseness* and *intent to harm* which are depicted in Figure 1.1.

Previous works have often focused on only one of these. For example, Conroy et al. (2015) focus on falseness/veracity, Rubin et al. (2015) on intentional deception and Wang (2017) on the deception that has intent to harm for some benefit, for instance, financial gain.

Wardle & Derakhshan (2017)'s definition for information disorder has been influential as evidenced, for example, by being used by The House of Commons (2018) using the term *disinformation* to refer to the fake news. The House of Commons (2018) adopted the definition of disinformation by Wardle & Derakhshan (2017) by defining *disinformation* as the deliberate creation and sharing of false and/or manipulated information that is intended to deceive and mislead audiences, either to cause harm or for political, personal or financial gain. The House of Commons (2018) defined *misinformation* as the inadvertent sharing of false information.

It is difficult for a human to segregate fake news from legitimate news. They are similar in the way they are written and shared in the social media. For a machine to detect deceptive contents in such homogeneity is a challenging task. Misinformation results from several interacting processes having five key elements: publishers, authors, articles, rumours and audience; which are equally responsible for its dissemination (Ruths 2019). One of the ways to detect deceiving content is by finding the stance (Mohammad et al. 2016*b*, Bourgonje et al. 2017, Zubiaga et al. 2018, Ghanem et al. 2018, Hanselowski et al. 2018). In this research I primarily focus on the stance detection.

1.1 Problem Statement and motivation

During a false online information debunking project *Emergent*, 1,660 articles were collected in the database within a time span from August to December 2014 amongst which 213 articles were identified with headline-body text pair dissonance (Silverman 2015). Sometimes a news body attached to a headline includes information and evidence that is disagreeing with the claim made in the headline. In the *Emergent* database, around thirteen percent of collected articles were found in which the news body disagreed with the headline (Silverman 2015). In other words for the thirteen percent of cases, the stance of body text does not match with the headline of the news as shown in the Example 1.1. There could be situations when articles are written in such a way that they discuss the claim made in the headline but, also include a hint of disapproval or doubt about the claim in the headline. Such hints (subtle or major) which oppose the headline, that could be known only after a reader goes meticulously through the complete article, are cause of misinformation.

Research (Gabielkov et al. 2016) shows that more than fifty percent of social media users share news headlines without clicking the link which contains the news article (DeMers 2016). In news reporting, a headline is generally a short text summarising an associated news article. Typically a reader expects the headline to be an accurate reflection of the article (Dor 2003). However, the existence of a corpus like FNC-1 by Pomerleau & Rao (2017), that is based on data originally collected by Silverman (2015), shows that the disagreement between news headline-body pairs is common enough to be a problem. Such disagreement can occur due to sensational headlines used for news reporting (Molek-Kozakowska 2013) or deliberately done to spread misinformation (Silverman 2015). News articles are often shared only by reading headlines if they agree with an existing belief system, known as confirmation bias, which may lead to the spread of misinformation (Knobloch-Westerwick & Kleinman 2012). Silverman (2015) found that in around thirteen percent of news articles, the headline does not reflect the content of the article's body.

The rapid spread of misinformation online is one of the highly prioritized issues (Forum 2014) that could show adverse effects in the future, if not tackled wisely today. When asked about fake news or misinformation, two-thirds of respondents said they encounter them at least once a week, and most respondents see it as a problem both in their country and for democracy in general. There were 1,500 respondents that include global experts across business, government, academia and civil society (Forum 2014). The existence of fake news is acknowledged as a serious issue by the public (European Commission 2018). These show how important it is to tackle the problem of online misinformation.

In my research, given a news article, I will focus on detecting useful text spans from the news body and will assess their usage in the stance detection. Those text spans are attribution relations. In addition I will do a fine-grained analysis of how different components of attribution relations that are **source**, cue and *content* contribute to the stance detection. The stance informs the readers whether the body text agrees, disagrees, discusses, or is unrelated to the headline. The additional stance information provides an aspect of validity of the news articles to the readers.

Previously, news article contents were used directly for stance classification (Yuxi Pan 2017, Andreas Hanselowski 2017, Benjamin Riedel 2017, Hanselowski et al. 2018, Ghanem et al. 2018, Slovikovskaya & Attardi 2020). This may not always be helpful because news articles also contain the author's opinion and additional contextual information. However, true information about any event or incident comes from sources that are mostly and explicitly mentioned in news. Thus, in my work I use such sources and their attributions rather than whole news articles to detect the stance.

1.2 Research Questions

My research question is

> **Can attribution relations and source expertise be useful in detecting the stance of a news article's body towards its headline?**

While reading news, can we believe information sources explicitly mentioned in news and messages communicated by those sources? The third-party source mentioned in the news by the authors and their respective claims is worthy to be explored. The main objective of this work is to provide a computational model that can highlight any disagreement between news headline-body pairs, also providing the source's credibility and respective claim information. Reich (2011) defined source credibility as "the degree to which the information from the source is perceived as accurate, fair, unbiased and trustworthy". In the following sub-section, I discuss the framework which I implemented to answer my research question.

1.2.1 Overview of my work: *AR-based 3C Framework*

The overview of my work is depicted in Figure 1.2. As we can see in the figure, there are three building blocks of the work that are grounded on the concept of the attribution relation(AR). I named my research framework in Figure 1.2 *AR based 3C framework*. An attribution relation is a way of attributing quotations to their respective speakers (He et al. 2013, Pareti et al. 2013, Newell, Cowlishaw & Man 2018). By identifying an attribution relation, three of its constituents are detected which are as follows (Pareti 2015, Newell, Cowlishaw & Man 2018).

- **source**: A communicative agent

- cue: Propositional attitude of the source. It is a lexical text that connects contents to their respective sources.
- *content*: Claims or propositions attributed to the source

The above-mentioned three different parts of an attribution relation could play different, but important roles to detect the stance of news bodies to their headline.

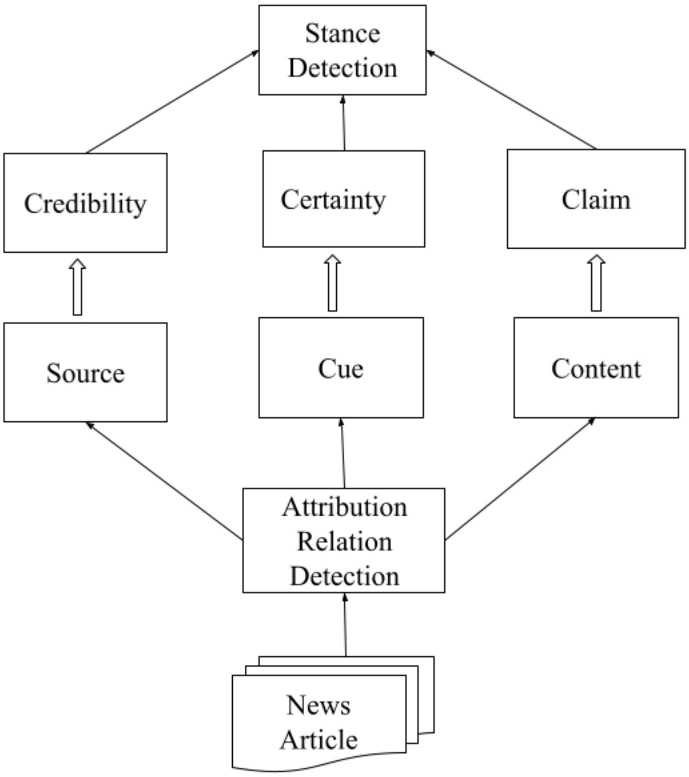


Figure 1.2: Overview of my work: *AR based 3C framework*

In my work, first I illustrate how we can detect attribution relations in news bodies. Then, I show the usefulness of those attribution relations in stance detection. Followings are three building blocks of my *AR based 3C framework* that deal with each part of an attribution relation.

- **Claim**: Attributed contents are referred to as *claims* in my work. I analyse how useful those claims are in the stance detection. Additionally, I use attribution relations instead of whole news articles as input for the stance detection. Then, I evaluate my work by comparing it with other state-of-the-art works.

- **Credibility**: I argue that an expert source provides credible information for news reporting. Source expertise is an attribute of source credibility. Thus, I analyse the role of source expertise in the stance detection by using attribution relations with only expert sources to an existing stance detection model. I evaluate its usefulness by analysing the stance detection model performance results with and without using attribution relations with non-expert sources.

- **Cue**: Attributional cues are those text spans which express a source's commitment or attitude towards its claims. I analyse the role of attributional cues in the stance detection by analysing whether removal of attributional cues degrades the stance detection performance.

As per Figure 1.2, we see that attribution relation is the central and novel idea used in my research for the stance detection. As AR comprises key information from the news body in the form of **source**, cue and *content* which are previously defined in this section, it can be directly used to detect the stance that is discussed later in Chapter 4. A news article can contain background details about an event and author's opinions as well. Detecting attribution relations can help us avoid such background details. Attribution relations allows us to represent a news article with only such information on the basis of which the news is written. Additionally, each AR component has many important features that contributes to the stance detection for instance, source's expertise. In my research, using attribution relation is a core aspect because it is a novel idea to use attribution relations in a news body instead of using the whole news body's content to detect the stance respective to the headline. In most of the works (Hanselowski et al. 2018, Slovikovskaya & Attardi 2020), whole news body is treated as a single entity to detect its stance to the headline with an exception (Sepúlveda-Torres et al. 2021) that used a summary of the news body instead of using whole news body content. In contrast, I opt to filter out background information and detect essential information in the form of attribution relation that is further used in the stance detection.

1.2.2 Research sub-questions

My research question is broken down to the following sub-questions.

RQ1: How can we detect attribution relations in a news article?

RQ2: Are attribution relations useful to detect the stance of a news body to its headline?

RQ3: Is a reader's judgement of claim credibility positively correlated with his/her judgement of the level of expertise of the source who is making that claim?

RQ4: What is the role of source expertise information in detecting the stance of a news body to its headline?

RQ5: Are attributional cues useful in the stance detection?

1.3 Research Contribution

There are five contributions of my research corresponding to each of the sub-questions discussed in Section 1.2 that are as follows:

C1: I present a new attribution detection model that can tag components of attribution relations in news texts. Additionally, I evaluated the model's performance on a different corpus.

C2: I treated attributed contents as potential claims from the sources. I illustrated that such claims are useful features to detect stance of a news article towards its headline. Additionally, I showed that attribution relations in a news article are useful bit of information to detect stance.

C3: My empirical study shows that in news articles a reader's perceived claim credibility is positively correlated with his/her perceived level of expertise of sources. I run statistical tests to show such correlation.

C4: I prepared an extended subset dataset of FNC-1 corpus. The FNC-1 corpus was introduced by Pomerleau & Rao (2017) in a fake news challenge and the corpus contains news articles with stance annotations. I collected source expertise data using a crowd-sourcing platform. Furthermore, I showed that source expertise data is useful in stance detection.

C5: I validated the usefulness of attributional cues to detect stance in news articles.

1.4 Thesis Outline

The introduction is followed by a literature review of stance detection in Chapter 2. I included literature reviews of the remaining works in their respective chapters. Chapter 3 corresponds to **RQ1** and **C1**. This chapter discusses a new model for attribution relation detection and its broader applicability in other domain data. In Chapter 4 that corresponds to **RQ2** and **C2**, I discuss the role of attribution relations in stance detection and show a comparative study with current best systems. Chapter 5 corresponds to **RQ3** and **C3** that explores the dependency of claim credibility on the source expertise for which I surveyed 25 participants. Following up on the conclusion from Chapter 5, in Chapter 6, I extended a subset corpus of FNC-1 with source expertise data that I collected in a crowdsourcing platform. Chapter 6 corresponds to **RQ4** and **C4**. In the same chapter, I also analyse the role of source expertise in stance detection. Chapter 7 corresponds to **RQ5** and **C5** that includes an assessment of attributional cues' role in stance detection. I did the final discussion about contributions and limitations in the last Chapter 8 with conclusions.

Chapter 2

Literature Review: Stance Detection

This chapter includes the literature reviews specifically related to the stance detection task. Literature reviews related to my sub-tasks of attribution relation detection, testing correlation of perceived claim credibility on the perceived source expertise, and usefulness of source expertise and cue features in stance detection are included in their respective Chapters 3, 5, 6 and 7. In this chapter, first I discuss the early works in the stance detection task and its relation to the rumour classification. I then discuss different corpora and shared challenges of stance detection focusing on news articles. I then discuss the evaluation metrics used in stance detection followed by the challenge imposed by headline-body length difference in news articles. Finally, we see different architectures used previously to detect stance. In each section of this chapter, I discuss the gaps and the rationale behind my choices.

Pomerleau & Rao (2017) defined stance detection as "an estimation of the relative perspective (or stance) of two pieces of text relative to a topic, claim or issue". I interpret stance detection as an automated task of estimating the stance of a piece of text (news body) to another piece of text (news headline), specifically in terms of the main claim conveyed by the headline. Stance detection is used as a foundation for several tasks such as assessing the veracity of micro-blog texts like Twitter (Aker et al. 2017, Derczynski et al. 2017), analysing online debates (Somasundaran & Wiebe 2010), understanding the argumentative structure of persuasive essays (Stab & Gurevych 2014) and assisting in detecting fake news (Pomerleau & Rao 2017). The stance detection problem is explored in the news domain (Ferreira & Vlachos 2016, Riedel et al. 2017) to assist the fact-checking process.

Stance can be used as a feature to decide the veracity of a claim because stance estimates the perspective of a given text in relation to that claim (Aker et al. 2017, Pomerleau & Rao 2017). Stance detection can also be useful to know what different news sources are publishing in relation to a certain claim or topic that can be a helpful step towards the veracity assessment of the claim (Pomerleau & Rao 2017). Stance detection is a crucial step to verify the truthfulness of a rumour (Zubiaga et al. 2018). In the next section, we see how the task of stance detection evolved and its relation with the rumour classification task.

2.1 Stance detection in rumour classification

A rumour is a circulating informational item which remains unverified until there is no evidence supporting it or no official confirmation from authoritative or credible sources (Aker et al. 2017, Zubiaga et al. 2018). In rumour classification, stance detection is defined as the classification of an author's attitude expressed in a text that can be supporting, denying, commenting on or querying a claim or fact (Aker et al. 2017, Zubiaga et al. 2018).

In rumour stance classification, the author's stance towards a claim is assessed whereas in my work we find the stance of a long text (news body) to another given text (headline), specifically in terms of the main claim conveyed by the headline. Rumour classification works (Aker et al. 2017, Zubiaga et al. 2016) used user and their claims as features to detect the stance that can be compared to two components- **source** and *content* of attribution relations. We can not know the propositional attitude of the user in rumours like Tweets that can be known in AR through cue. Thus, we can say that my work has partial resemblance to rumour classification on the basis of information used to detect the stance. Furthermore if we consider the rumour stance classification in the most frequently used Twitter data, the text length difference between two pieces of texts (a rumour and a user's response) is much lesser than the news headline-body pairs in my work. Tweet length is limited to 280 characters (140 characters before 2017). In contrast, the length of a news body is not restricted to a certain number of characters. Thus, higher discrepancy between news headline-body pairs in my work pose challenges in the stance detection.

The current work, however, focuses on longer texts, such as news articles. Thus, unlike stance detection in rumours, my computational model for stance detection aims to automatically find the orientation of news bodies to their associated headlines. Such stance classes can be any one of the following:

- agree: The body text agrees with the headline.
- disagree: The body text disagrees with the headline.
- discuss: The body text discusses the same topic as the headline, but does not take a position.
- unrelated: The body text discusses a different topic than the headline.

Early work in rumour stance detection by Qazvinian et al. (2011) classified a user's belief in a rumour as believe or deny/doubtful/neutral. They used Twitter data (short text with a maximum character limit of 140) for two-way belief classification. Another work by Lukasik et al. (2015) in contrast to Qazvinian et al. (2011)'s work used three classes- *support*, *deny* and *question* to classify tweet-level judgement of newly emerging rumours. Lukasik et al. (2016) introduced the fourth class of rumour stance called *commenting*. The fourth class was introduced to remove noise from *supporting* and *denying* classes. However, the class *commenting* has no contribution to the veracity assessment of the rumour. Zubiaga et al. (2016) mentioned that rumour stance classification can assist in verification tasks by aggregating the stance of multiple tweets discussing a rumour or by deriving a consensus from the stance of what Twitter users reply to each other related to a rumour. Similar work by Aker et al. (2017) argued that for veracity assessment where a claim is already known, people's reactions can be gathered and observed to decide the truthfulness of the claim. Zubiaga et al. (2018) introduced a pipeline framework for rumour classification with four components- rumour detection, rumour tracking, stance classification and rumour veracity classification as shown in Figure 2.1.

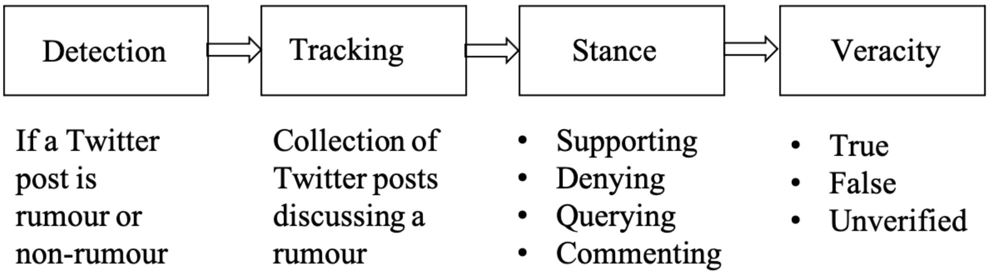


Figure 2.1: The pipeline framework for rumour classification adapted from Zubiaga et al. (2018)

As shown in Figure 2.1, stance detection is a step towards the veracity classification of a rumour. Conforti et al. (2018) highlighted steps in the rumour classification pipeline relative to data in the FNC-1 corpus for the fake news detection. Conforti et al. (2018) argued that the tracking phase deals with filtering and selecting only those contents that are related to the considered topic. The stance detection phase determines if there exists any information deviation between the news headline and its associated body. Conforti et al. (2018) discarded the unrelated class of the FNC-1 corpus to detect the stance because they consider that the unrelated class data belongs to the tracking phase. The decision by Conforti et al. (2018) to exclude the unrelated class from the stance detection task looks appropriate considering the unrealistic nature of the unrelated class that has topically different news headline-body pairs combined to form a news article. Dropping noisy and frequently appearing unrelated data from the computation can enable the system to focus on the remaining minority classes. However, I observed that Conforti et al. (2018) is the only published work that excluded unrelated class data, which makes it ineffective to compare it with current best systems. Moreover, systems by Zhang et al. (2019), Slovikovskaya & Attardi (2020), Sepúlveda-Torres et al. (2021), which consider all four stance classes, showed better stance detection performance than the work by Conforti et al. (2018). So, I decided to use the following stance classes in my work that are previously defined in this section:

- agree
- disagree
- discuss
- unrelated

We can compare the given stance classes with the ones in the rumour classification shown in Figure 2.1. The classes agree and disagree corresponds to the rumour stance labels *supporting* and *denying* respectively as they show a clear support and opposition to the target claim or rumour without any hedging. In rumour stance classification, *querying* refers to such posts that asks questions or appeals for more information about a rumour. Furthermore, the label *commenting* refers to a post which contains comment that does not contribute in any way to the veracity of the rumour. I believe the discuss class data may include queries as well as comments about the target claim. Thus, discuss class contains texts that can be similar to rumour stance classes *querying* and *commenting*. Rumour stance classes do not include the unrelated class because contents that are not related to a rumour are filtered out during the tracking phase of Figure 2.1 such that only those posts or texts are selected that discuss a rumour. Table 2.1 shows examples of four stance classes from the FNC-1 corpus that I and several other works (Zhang et al. 2019, Slovikovskaya & Attardi 2020, Sepúlveda-Torres et al. 2021) used for the stance detection. Table 2.1 includes stance classes with their corresponding headline-body pairs of news. Conforti et al. (2018) did not use the unrelated class considering it belonging to the tracking phase of Figure 2.1.

Headline	Student accidentally sets college on fire during fireworks proposal	Spider burrowed through tourist's stomach and up into his chest	Mystery of 50ft giant crab caught on camera in Kent harbour	Christian Bale in talks to play Steve Jobs for Sony, Danny Boyle
News body extract	The plan bombed, though, when the fireworks set the grass ablaze at the Liaoning Advertisement Vocational College in the city of Shenyang. As firefighters rushed extinguish the massive blaze, Chien, searched for his girlfriend, who forgot that he had asked her to join him for a walk.	Arachnologist Dr Volker Framenau said whatever the creature was, it was "almost impossible" for the culprit to have been a spider. "If you look at a spider, the fangs, the mouth parts they have, they are not able to burrow. They can't get through skin," he said.	The photograph was posted on a website called Weird Whitstable - an online collection of strange and unusual sightings in the town. Its curator, Quinton Winter, said that at first he thought the image - sent to him by a follower - showed an unusual sand formation, but that he is now convinced it is a monster of the deep.	Apple crushed its introduction of the Apple Watch yesterday in Cupertino, but while Kevin Lynch and Jony were waxing poetic about the design of watch and its revolutionary UI, there was one feature everyone steered clear of: battery life.
Stance	agree	disagree	discuss	unrelated

Table 2.1: Examples of stance classes from the FNC-1 corpus (Pomerleau & Rao 2017)

2.2 Resources and shared challenges

Stance detection is mostly explored in short-length texts like Twitter data (Zubiaga et al. 2016, Mohammad et al. 2016*a*,*b*, Derczynski et al. 2017). Ferreira & Vlachos (2016)'s work is the first to include news documents (long texts) in their stance classification corpus. However, Ferreira & Vlachos (2016) did not use news bodies' contents in their work. They detect the stance of news headlines relative to a specific claim and this stance can be *for*, *against* or *observing*. For example,

Claim	Two Australian men kept a McDonald's Quarter Pounder with cheese for 20 years	Artist Banksy drew a cartoon about the Charlie Hebdo murders	The Batmobile was stolen
Headline	20 year old burger? McDonald's burger purchased in 1995 hasn't aged a bit	Banksy's illustrated response to Charlie Hebdo attack isn't by Banksy. But it is striking	It's back! The incredible new Batmobile takes a spin around Detroit.. after rumors it had been STOLEN
Stance	*for*	*against*	*observing*

Table 2.2: Examples from the Emergent Corpus by Ferreira & Vlachos (2016)

Ferreira & Vlachos (2016) used claims and news documents collected during the Emergent project by Silverman (2015). Pomerleau & Rao (2017) extended the corpus by Ferreira & Vlachos (2016) with an addition of document level stance and named it FNC-1[1]. Pomerleau & Rao (2017) introduced the FNC-1 corpus in a fake news challenge (FNC) to detect stance of news bodies relative to their headlines that can be agree, disagree, discuss or unrelated. The objective of the challenge was to develop AI-assisted fact-checking techniques to combat fake news. As I was interested to work with long texts like news articles, I opted to use the FNC-1 corpus for the stance detection task which is a helpful step toward identifying fake news.

[1] https://github.com/FakeNewsChallenge/fnc-1

To assist in capturing information disorder, there are many shared challenges of stance detection in short texts like Twitter data (Mohammad et al. 2016*b*, Derczynski et al. 2017). However, so far to my knowledge Pomerleau & Rao (2017)'s FNC is the only one which covered document-level stance detection intended to assist fake news detection. Rubin et al. (2015) defined fake news detection as the prediction of the chances of a particular news article being intentionally deceptive. From this definition and according to definitions by Wardle & Derakhshan (2017) discussed in Chapter 1, we can say that fake news is disinformation.

The FNC-1 corpus consists of 75,385 news articles that is divided into training and test sets with 49,972 and 25,413 news articles respectively. The FNC-1 corpus has four stance classes- agree, disagree, discuss and unrelated (Pomerleau & Rao 2017). The FNC-1 corpus contains higher percentage of unrelated class with 54,894 data and rest distributed to remaining three classes. The percentage data distribution for agree, disagree, discuss and unrelated classes are 7.4%, 2.0%, 17.7% and 72.8% respectively. As the unrelated class contains topically different headlines and news bodies associated together, for the system evaluation the FNC awarded 0.25 for such classification considering it an easy prediction task. Correct predictions to the remaining three classes are given 0.75 additional points.

The FNC organisers provided a baseline model that used hand-coded features and a Gradient Boosting classifier.[2] The hand-coded features included word/n-gram overlap features, and polarity and refutation indicating features. The baseline achieved a weighted accuracy score of 79.53%. Fifty teams participated in the FNC using varieties of techniques involving statistical methods, neural methods and hand-crafted features. The winner of the challenge implemented a combination of deep convolutional neural networks and gradient-boosted decision trees with word embedding and lexical features.[3] The winner got weighted accuracy score of 82.02%. Although the FNC adopted weighted accuracy as the evaluation metric, many

[2] `https://github.com/FakeNewsChallenge/fnc-1-baseline`

[3] `https://github.com/Cisco-Talos/fnc-1`

works preferred other metrics that I will discuss in the next section.

2.3 Evaluation metrics

Despite using a weighted accuracy metric (Pomerleau & Rao 2017) for the performance evaluation of the stance detection system, Hanselowski et al. (2018) argued that the accuracy is not an appropriate metric to evaluate the highly imbalanced FNC-1 corpus. Hanselowski et al. (2018) proposed macro-averaged F-score metric for the system performance evaluation. The class-wise evaluation of the stance detection system is useful to show how well the system performed for each class labelling. It helps us to focus on rare classes like disagree and agree that are of my prime concern. Hanselowski et al. (2018) illustrated that even the top winning systems of the FNC have very poor results for class-wise stance predictions of the disagree class. Such poor performance was not reflected by the metric used in the FNC. The disagree class is the class of our prime concern because it consists of news headlines that do not reflect the contents of their respective news bodies. The evaluation metric used by Hanselowski et al. (2018) is widely adopted (Slovikovskaya & Attardi 2020, Sepúlveda-Torres et al. 2021, Roy et al. 2022) to evaluate the stance detection system's performance in the FNC-1 corpus.

Considering the imbalanced distribution of data in the FNC-1 corpus and the aforementioned works, I decided to evaluate the class-wise performance of my stance detection system along with a macro-averaged F-score.

2.4 Challenging news headline-body length difference

Besides the imbalanced data distribution in the FNC-1 corpus (Pomerleau & Rao 2017), another challenging factor of the FNC-1 corpus is the text length difference between the news headline and the associated body. In the FNC-1 corpus, the longest headline length is 40 words and the longest news body length is 4788 words. Several other works (Ghanem et al. 2018, Conforti et al. 2018, Sepúlveda-Torres et al. 2021) highlighted computational complexity in predicting a long text's (news body) stance to a short length text (headline). Short headlines with no claims like *Crabzilla*, *Bali Awry*, *Staff Reporter*, *Ghost Ship* in the FNC-1 corpus are tricky to handle. Even for a human reader, it is difficult to decide whether the news body is agreeing or disagreeing with such a headline. I argue that there should be an explicit claim in the headline for the stance detection task like in the Example 1.1. Without any claims in the headline, it is difficult to decide whether the news body is agreeing or disagreeing with the headline. I can't say how Pomerleau & Rao (2017) did the stance labelling in that context because they did not provide any overview paper or a detail annotation guideline for the FNC-1 corpus.

A work by Ghanem et al. (2018) highlighted that it is difficult to handle a high discrepancy between lengths of headlines and new bodies while detecting the stance of a news body to its headline. Another work by Conforti et al. (2018) presented highly uneven lengths of headlines and news bodies as a key challenging feature. Conforti et al. (2018) argued that the most important information resides at the beginning of a news body that can be useful to classify the stance relative to the headline. To deal with such length differences, I propose a model to extract attribution relations from news bodies. Representing a news body by attribution relations not only reduces the length of the news body by excluding the author's opinions and additional information related to the news; but also represents the news by sources and their claims based on whom the news is written. So far to my knowledge, my work is the first to explore the usefulness of attribution relations to detect the stance of news bodies relative to their headlines.

Most works in stance detection (Hanselowski et al. 2018, Ghanem et al. 2018, Zhang et al. 2019, Slovikovskaya & Attardi 2020) used whole news body contents except the work by Sepúlveda-Torres et al. (2021) which used a summary of the news body to label the stance of a news body to its headline. Sepúlveda-Torres et al. (2021) argued that the news headline-body asymmetry can be resolved by representing the news body by its summary. Sepúlveda-Torres et al. (2021) compared a news body's summary to the headline to classify the stance instead of using a whole news document. My work is similar to their work, with the main difference that I represent a news body by attribution relations in it that are further used in the stance detection.

2.5 System Architecture

With the introduction of the FNC-1 corpus in the fake news challenge held in 2017, different works used a variety of techniques that involved machine learning methods (Hanselowski et al. 2018, Slovikovskaya & Attardi 2020), word embedding (Yuxi Pan 2017, Ghanem et al. 2018) and hand-crafted features (Andreas Hanselowski 2017, Benjamin Riedel 2017). The top winning team of the challenge implemented a combination of convolutional neural network (CNN) and gradient-boosted decision tree (Tree) models (Yuxi Pan 2017). The CNN model used word2Vec for word embedding whereas the Tree model implemented features like word count, TF-IDF, sentiment, and singular-value decomposition in combination with word embedding. The second placed system in the FNC by Andreas Hanselowski (2017) used a multi-layer perceptron with a variety of hand-engineered features. The features include uni-gram, cosine similarity of word embeddings of nouns and verbs between headline and document tokens and topic models based on non-negative matrix factorization, Latent Dirichlet Allocation, and latent semantic indexing in addition to the baseline features provided by the FNC-1 organizers (Andreas Hanselowski 2017). The third placed team in the challenge implemented a multi-layer perceptron with features like term frequency vectors of unigrams of the 5,000 most frequent words for the headlines and the documents concatenated with the cosine similarity between the TF-IDF vectors of the headline and document (Benjamin Riedel 2017).

The second placed team of the FNC came with new work in the same corpus. The work by Hanselowski et al. (2018) argued that stacking layers adds more representational power to a neural network. Hanselowski et al. (2018) implemented a stacked bi-directional long short-term memory network with word embedding. Their system had state-of-the-art results while evaluating the system's performance using a macro-average F-1 score. Hanselowski et al. (2018) proposed macro-average F-1 score as a suitable metric for the highly imbalanced FNC-1 corpus. Their work also showed that despite having high average accuracy, the winning system of the FNC had a very poor performance for the disagree class. Thus, to reflect the class-wise performance of the system in imbalanced data distribution, a macro-average F-1 score is an appropriate choice for the system evaluation.

Another work in the FNC-1 corpus by Ghanem et al. (2018) implemented a neural network architecture with two hidden layers with a rectified linear unit (ReLU) activation function as non-linearity for the hidden layers and a Softmax activation function for the output layer. Ghanem et al. (2018) used cosine similarity between the embedding of each sentence for each headline-body tuple, hand-curated cue features and baseline features used in the FNC. Ghanem et al. (2018) also mentioned that a neural network architecture worked better than other machine learning models like Support Vector Machines (SVM), Gradient Boost, Random Forest and Naive Bayes classifiers in their setting.

In contrast to previous works in stance detection where stance detection was considered a classification problem, Zhang et al. (2018) proposed a ranking-based method. Zhang et al. (2018) used a multi-layer perceptron with two hidden layers that produce a value for true stance and three values for wrong stances. Ranking loss functions in those hidden layers maximize the value difference between the true and false stances. Zhang et al. (2018) showed that their system is more effective than the top-three winning systems of the fake news challenge.

So far in most of the works, stance detection has been handled as a four-way multi-classification problem. Works like (Hanselowski et al. 2018, Slovikovskaya & Attardi 2020) handled the stance detection problem as a single stage problem where the output can be any one of four given labels agree, disagree, discuss or unrelated. The stance detection model in such cases can be highly biased towards most frequently occurring unrelated class resulting in a poor performance for the least occurring disagree class. The shortage of representative samples in a class degrades the model performance for that class. However, Zhang et al. (2019) argued that a hierarchical architecture is useful to mitigate the class imbalance problem in the FNC-1 corpus rather than doing four-way multi-classification as done in previous works. The hierarchical architecture by Zhang et al. (2019) implemented a two-layered neural network where the first layer segregates the unrelated class from the rest of the related classes. At the second layer, related data are classified as agree, disagree or discuss classes. I used the FNC-1 corpus (Pomerleau & Rao 2017) that is unevenly distributed with around 73% of data in the unrelated class. Other remaining classes (agree, disagree and discuss) are the subject of my interest because I believe those samples are more frequently seen in the real scenario. The unrelated class contains news articles with topically different headlines and news bodies. Zhang et al. (2019) used a two layered neural network where only at the second layer classification among minority classes like agree, disagree and discuss takes place. The highly occurring unrelated data is handled at the first layer of the neural network.

Following their work, I opt to implement two-stage system architecture in my work. At the first stage, I filter the unrelated class data, which contains topically different news headlines and bodies associated to form a new document, from the rest of the classes. In the second stage, the stance of a news body to its headline is classified as any one of three given labels agree, disagree or discuss. Slovikovskaya & Attardi (2020) argued that the stance detection task can benefit from transfer learning in pre-trained transformers and showed better stance detection performance in the FNC-1 corpus than previous works. Following their work, I decided to use Simple Transformers by Rajapakse (2017) that are built on top of Hugging Face transformers to implement pre-trained models in my stance detection system.

Similar to Zhang et al. (2019)'s hierarchical architecture, Sepúlveda-Torres et al. (2021) proposed a two-stage classification architecture for stance detection in the FNC-1 corpus. The first stage is named *Relatedness Stage* and the second is called *Stance stage*. In the *Relatedness Stage*, Sepúlveda-Torres et al. (2021) performed a binary classification where each headline-body pairs are classified as *related* or *unrelated*. The second *Stance stage* involves three-way multi-class classification where each headline-body pairs are classified as *agree*, *disagree* or *discuss*. Sepúlveda-Torres et al. (2021) implemented hand-engineered similarity and polarity features at the first and second stages respectively. Similar to Slovikovskaya & Attardi (2020)'s work, Sepúlveda-Torres et al. (2021) used a transformer-based RoBERTa model for classification in both stages. The novel approach in Sepúlveda-Torres et al. (2021)'s work is the use of a summary instead of a whole news body in the stance detection task. Sepúlveda-Torres et al. (2021) used an off-the-shelf summary extraction system to extract a summary containing the most relevant five sentences from the news body. My approach to stance detection is similar to Sepúlveda-Torres et al. (2021)'s work with the distinction that I am representing a news body by attribution relation present in it. An attribution relation has information like who the source is, what claim the source made and how much commitment the source is making to the claim. Additionally, I build my system for such component detection that are parts of an attribution relation.

Recent work by Roy et al. (2022) implemented a three-stage approach with problem-specific features in each stage. In the first stage, they filtered unrelated data from the rest. Secondly, they filtered neutral news articles labelled as discuss. Finally, at the third stage Roy et al. (2022) implemented an agree/disagree binary classifier. Their machine learning models with task-specific features outperformed the performance of Hanselowski et al. (2018) by 1%, however Roy et al. (2022) didn't outperform current state-of-the-art results. Furthermore, Roy et al. (2022) doesn't include their system performance comparison with the current best systems. However, the three-stage approach proposed by Roy et al. (2022) seems useful if we employ content-specific features in each stage of the stance detection.

2.6 Conclusion

Despite Conforti et al.'s (2018) argument of removing unrelated class articles considering those irrelevant to the stance detection, I opt to use all four stance labels for the comparative analysis with the state-of-the-art systems. Furthermore, influenced by Zhang et al. (2019) and Sepúlveda-Torres et al. (2021)'s system architectures, I employed a two-stage approach where the first stage deals with filtering out unrelated data and the second stage with the actual stance detection task. I represent news articles by attribution relations that include sources, their attributed claims and the source's propositional attitude. Using attribution relations helps to accommodate the difference between a headline and body lengths, but without losing essential information about the body. Moreover, I decided to evaluate the stance detection performance using class-wise and macro-averaged F-score as proposed by Hanselowski et al. (2018) to address the imbalanced data distribution in the FNC-1 corpus.

Chapter 3

A new model for Attribution Relation Detection

3.1 Introduction

In this section, I describe a new model for attribution relation (AR) detection. The attribution relation detection model, which I also call "AR model" in short throughout this dissertation, is used to extract attribution relations from news articles written in the English language. The objective of identifying an attribution relation consists of finding sources with their respective claims along with the source's propositional attitude. The attribution relation can be used to help assess the trustworthiness of a news article and to find the news article's stance towards the headline. Here, I use attribution relations in a news body to predict its stance to the associated news headline. In this chapter, I intend to answer the following research question:

> RQ1: How can we detect attribution relations in a news article?

I intend to handle the research question RQ1 by building a new model which detects different components of an attribution relation in a news body that includes sources, their respective claims and propositional attitudes. I call it a new model because in previous works for attribution detection (Pareti 2015, Newell, Cowlishaw & Man 2018), a pipeline of classifiers (such as knearest neighbour algorithm, conditional random field network) is used to detect each component of an attribution relation. Additionally, Pareti (2015), Newell, Cowlishaw & Man (2018) used different models to link source, cue and content to get an attribution relation. In contrast, I implement a deep learning model which tags all components of an AR in a single phase. My model can tag all AR components in a sentence at the same step. This helps us get rid of requiring separate models for each AR component detection in the news articles. Attribution relations are relational texts that should be detected in a given order. Thus, for its better prediction a model with some buffering or memory can be useful that can track the relation among the components. So, I opt to use Bi-directional long short-term memory (Bi-LSTM) to detect attribution relations that showed promising performance in other sequence labelling tasks. Additionally, in similar NLP works like named entity recognition, Bi-LSTM network showed better performance than conditional random field network.[1] This contrasts my single-step approach for AR detection with separate components of verb-cue classifier, source and content classification models by Pareti (2015).

Attribution relation is a text combination where we attribute an object such as a piece of text to its respective source or speaker through a lexical cue. In other words, the relation which binds some quotes or propositions to their respective source via a connective is called an attribution relation. Following the definitions of attribution relations by Pareti (2015), Newell, Cowlishaw & Man (2018), Newell, Margolin & Ruths (2018), I define three components of an attribution relation as:

- Source: A speaker or some report to which claims are attributed.
- Cue: A lexical anchor that expresses the source's knowledge, attitude or intention towards someone or something. It is a reporting phrase that includes associated auxiliaries and negations along with the main verb.
- Content: A part of text that is attributed to the source. In my work, it

[1] https://github.com/moejoe95/crf-vs-rnn-ner

can also be considered as potential claims from different sources.

For the detection of attribution relations in a text, I implemented a neural network model with two bi-directional long short-term memory networks and Embeddings from Language Models (ELMO) (Peters et al. 2018) for word embedding. The input to the model is a sentence given as a list of tokens. The output of the model is any of the four given tags for each token- **source**, cue, *content* or O. For example,

Input:	**The**	**boy**	told	News	Corp	*that*	*it*	*was*	*a*	*spider*	.
Output:	**source**	**source**	cue	O	O	*content*	*content*	*content*	*content*	*content*	O

I represent the three components of an attribution relation using specific fonts as **source**, cue and *content* in this dissertation. The tag O represents that the token is not a part of an attribution relation. The following example in Figure 3.1 shows how each token in an input sentence is processed to its respective label as the output.

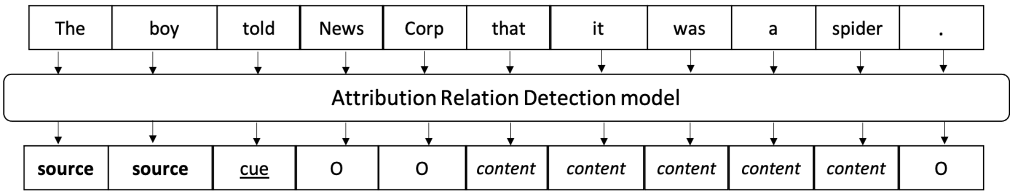


Figure 3.1: Attribution Relation Detection model

3.2 Literature Review

This section describes research works related to attribution relation detection including history, different terminologies and annotation schemes. Moreover, I discuss different corpora of attribution relations and methods used for their detection.

Within the classical framework of Rhetorical Structure Theory (Mann & Thompson 1988), Marcu (1999) and Carlson & Marcu (2001) presented attribution as a rhetoric relation with a satellite containing a source and an attribution verb of speech or cognitive act. Marcu (1999) and Carlson & Marcu (2001) presented an attributed message as the nucleus. Further, Prasad et al. (2006) defined attribution in discourse relations as a relation of ownership between an individual and its respective attributional objects. In Penn Discourse Tree Bank (PDTB) corpus, a discourse relation is characterized by four different features (Prasad et al. 2006):

Source Different agents.

Type Nature of relationship between agent's attributed contents.

Scopal polarity Marks the negation which reverses the polarity of attributed content.

Determinacy Captures if an argument can itself be cancelled in particular contexts, such as within the scope of negations, conditionals, or infinitivals.

The following is an example from PDTB corpus by Prasad et al. (2006) where discourse connective is explicit and underlined. Argument1 and Argument2 are *italicised* and **boldfaced** respectively. Here, the source as "Ot" represents that the speaker is someone other than the author; and "Inh" indicates that the source value is inherited from the relation. Type as "Comm" refers to assertions and "Patt" represents a propositional attitude. "Neg" refers to the presence of scopal polarity. "Null" represents the absence of the feature.

Example 3.1.
"Having the dividend increases is a supportive element in the market outlook, but I don't think **it's a main consideration,"** he says.

	Discourse Rel	**Arg 1**	**Arg 2**
Source	Ot	Inh	Inh
Type	Comm	Null	Patt
Polarity	Null	Null	Neg

We can see attribution in discourse relations as a fundamental concept for attribution relations. They are further conceptualized in different ways.

The concept of attribution relation is similar to that of quote attribution. However, many works in quote attribution focused to attribute direct quotes, in areas such as literary texts (Elson & McKeown 2010, He et al. 2013, Yeung & Lee 2017), childrens' stories (Iosif & Mishra 2014) and news articles (Pouliquen et al. 2007), using either rule-based approaches or sequence labelling (O'Keefe et al. 2012). An example of quote attribution from the literary narrative that has utterances attributed to a character is as follows (He et al. 2013):

> *As they went downstairs together, Charlotte said, "I shall depend on hearing from you very often, Eliza."*

In the given text, *"I shall depend on hearing from you very often, Eliza."* is the quote which is attributed to the speaker *Charlotte*. Attributed utterances may not always be direct, they can be indirect or mixed. Thus, detecting utterances within inverted commas is not enough to find all utterances in a document. My work deals with detection of both direct and indirect quotes.

The concept of quote attribution is extended by Pareti (2012) to attribution relation by including all components involved in a quote attribution that are the source making the utterance, the source's propositional attitude while making the utterance and the utterance itself. Pareti (2012) used the PDTB corpus by Prasad et al. (2006), and defined attribution relation as a composition of three fundamental elements- source, cue, and content span. Pareti (2012) included an additional supplementary span which include information that is relevant to the *content* such as location, date, time and so on. The corpus with annotation of such attribution relations is termed the Penn Attribution Relation Corpus (PARC). The following is an example from the PARC corpus where an attribution relation is represented by a combination of **source**, <u>cue</u> and *content*:

> **The assistant HHS secretary** <u>said</u> *the ban "should be continued indefinitely."*

Although a supplement span contains useful information, it is not an essential element of an attribution relation (Pareti 2015), so I don't consider it as a component of an attribution relation in my work. For example,

> **He** told News Corp yesterday *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin*.

Here, "News Corp yesterday" is a supplement span with two parts "News Corp" and "yesterday" that respectively represent the recipient of an assertion and when the assertion was made. The corpus by (Newell, Margolin & Ruths 2018) that I will use for attribution relation detection does not have supplement spans annotated.

Attribution relations may exist as any of direct, indirect or mixed quotes. Attribution relations not only express speech acts but, also intention, knowledge or belief of sources (Pareti 2015). Pareti (2015), Newell, Cowlishaw & Man (2018) implemented pipelines of different classifiers to identify AR components separately and resolve relationships to extract an attribution relation. In contrast, I implemented a deep learning model to detect attribution relation that tags each token in a sentence as **source**, cue or *content*. There is no pipeline in my work to handle each component separately. My model handles attribution detection as a sequence labelling task that I will explain later in Section 3.6.

Source-cue-content triplet identification as attribution relation by Pareti (2015) has strict results of precision, recall and F-score as 63%, 50% and 56% respectively on the PARC corpus. Here, strict results refer to those attribution relations that have all their constituents correctly predicted. Table 3.1 shows strict classification performance of three different classifiers that are responsible for detecting source, cue and content of an attribution relation.

Classifier	Precision	Recall	F-score
Verb-cue	0.90	0.90	0.90
Content	0.80	0.64	0.71
Source	0.84	0.84	0.84

Table 3.1: Strict performance of verb-cue, content and source classifiers (Pareti 2015)

Following the work by Pareti (2015), Newell, Cowlishaw & Man (2018) made a slight improvement on the overall performance of the quote extraction system with precision, recall and F-score as 62.1%, 52.2% and 56.8% respectively. The overall performance was measured using the intersection of the PARC and CoNLL-2011 corpora with an exact match for all quote spans after resolving coreferences. Newell, Cowlishaw & Man (2018) used a pipeline approach similar as Pareti (2015) to detect the source, cue and content spans separately that has a better performance as given in Table 3.2.

Classifier	Precision	Recall	F-score
Verb-cue	0.97	0.85	0.91
Content	0.76	0.68	0.72
Source	0.92	0.89	0.91

Table 3.2: Performance of verb-cue, content and source classifiers (Newell, Cowlishaw & Man 2018)

Table 3.2 illustrates that their system worked better for source and verb-cue classification than content labelling. Generally, source and cue spans are shorter than content spans. A less efficient computational model for the content classification means losing more claims, which further implies losing more attribution relations. Although these good works might have been useful to evaluate a new system, the PARC corpus is not freely available. Therefore, I used an openly available corpus containing Political news documents by Newell, Margolin & Ruths (2018).

3.3 Data

We used an openly available corpus of attribution relations called Political News Attribution Relations Corpus (PolNeAR) by Newell, Margolin & Ruths (2018). The articles in PolNeAR are taken from 7 US national news publishers- Huffington Post, USA Today, Western Journalism, Washington Post, Politico, Breitbart and New York Times. The news articles are related to the campaigns of the US General Election 2016. Newell, Margolin & Ruths (2018) included only such articles that mention at least one of the candidates, Donald Trump or Hillary Clinton. Newell, Margolin & Ruths (2018) collected news articles uniformly by selecting 84 articles per month for 12 months period between 8 Nov 2015 to 8 Nov 2016.

For the corpus validation, 6 trained annotators manually annotated 36% of randomly selected data from the PolNeAR corpus. The inter-annotator agreement for attribution annotation in the PolNeAR corpus is expressed by the *agr* metric (Wiebe et al. 2005) and Krippendorf's alpha (Krippendorff 2018). The PolNeAR corpus has a high *agr* metric of 92.3% and Krippendorf's alpha of 0.754. In the PolNeAR corpus, the *agr* metric measures the extent to which annotators agree on the existence of an attribution relation without any concern for the boundary. For example, two annotators are agreeing if one of them annotated a source as *Donald Trump* and another as *Donald Trump, the President of the U.S*. The *agr* metric measures the raw percentage agreement. For Krippendorf's alpha, each token is treated as a separate labelling decision. It means annotators have to agree at the token-wise level. The inter-annotator agreement (*agr* metric) in the PolNeAR corpus is higher than that in the PARC3 corpus (Pareti 2016) by 9%. It implies that attribution relation data in the PolNeAR corpus is more reliable than that of the PARC3 corpus.

PARC3 by (Pareti 2016) is an often used corpus to computationally detect attribution relations in news texts. An earlier version of the corpus PARC2 by Pareti (2012) was not fully annotated. In PARC2 copus, a proportion of 30-50% of attribution relations are unlabelled (Pareti et al. 2013). Newell, Margolin & Ruths (2018) argued that PARC3 corpus has majority attribution relations unlabelled that they conclude by re-annotating 56 randomly selected articles form the PARC3. Extrapolating the rate of unlabelled ARs to the full corpus, Newell, Margolin & Ruths (2018) argued that more than 20 thousand attribution relations are missing in PARC3. Considering the larger size and completeness of the corpus, I decided to use the PolNeAR corpus by Newell, Margolin & Ruths (2018) in my work of computational detection of ARs.

The PolNeAR corpus is divided into three subsets of data: training set, development set and test set. The development set of data is termed the validation set in my work. Data distribution over multiple sets are shown in Table 3.3.

	Training	**Validation**	**Test**
News articles	840	84	84
Attributions	19865	2191	2047
Sentences	30232	3308	3022
Sentences with AR	21231	2374	2150

Table 3.3: PolNeAR corpus statistics

Training and validation sets of data are respectively used to train and validate the AR model. The unseen test set is utilised to check how well the AR model performs on unseen data. In the next section, I discuss what input I give to the AR model and what output I expect from it.

3.4 Input and Expected output

The input to the AR model is a news body that is represented as a list of sentences. Sentences are further represented as lists of tokens. The PolNeAR corpus by Newell, Margolin & Ruths (2018) processed news articles with Stanford's CoreNLP software (Manning et al. 2014) to provide tokenization, sentence splitting, POS tagging, constituency and dependency parsing, named entity recognition, and coreference resolution. I used sentence splitting and tokenization features among those to find a list of tokens in each sentence of a news body.

A list of tokens in a sentence is fed to the AR model as the input. We can use a sentence as input with up to 50 tokens. If it is longer than this, then exceeded tokens are discarded. The expected outputs are the tags for each token in the sentence. Each token in the sentence is labelled as one of four given tags: **source**, cue, *content* or O. The tag O represents that the token is not a part of an attribution relation.

Example 3.2.

Input: O'Malley 's campaign says it will shift staff from its Baltimore headquarters to the early states .

> **Output**: **source source source** cue *content content content content content content content content content content content content* O

Once we have a system that can give the expected output when an appropriate input is given to it, we have to evaluate the overall performance of the system. To evaluate how well my AR model works for AR detection, I need to set a baseline that is discussed in detail in the next section.

3.5 Baseline

To evaluate how well my model works for the AR detection, I need a baseline for the comparison. I implemented the majority class baseline. Many research works in NLP used the majority class baseline as the sole baseline especially when there is an unavailability of similar work for the comparison. For examples, works by Somasundaran & Wiebe (2010) and Stab & Gurevych (2014) used the majority class baseline in their respective works of classifying stance in ideological debates and finding stance to identify the structure of argumentative discourse.

Furthermore, I chose the majority class baseline because of the following:

i. To the best of my knowledge, there is not any attribution relation detection system available that used the PolNeAR corpus.

ii. Other work in attribution relation detection like Newell, Cowlishaw & Man (2018) is not openly available, even not through personal communication.

In the majority class baseline, a classifier simply labels every instance of the input with the majority tag for the corresponding target. Thus, all tokens in the input are tagged as the label that is the most frequent attribution label in the corpus. According to the Table 3.4 that shows attribution label distribution in different datasets of the PolNeAR corpus, *content* is the highest occurring label. So, I consider that all tokens are predicted as *content*. As per Table 3.4, the token label *content* occurs more often than **source**, cue, O in all training, validation and test sets of the PolNeAR corpus.

	Training	Validation	Test
source	56274	6019	5684
cue	37770	3995	3745
content	340457	39047	36467
O	290147	29287	27474

Table 3.4: Token label distribution in the PolNeAR corpus

3.5.1 Token-wise Baseline

While considering token-wise labelling of attribution relations, we get results as shown in Table 3.5. The table shows the token-wise baseline in the test set of data. The table shows the model performance to label each token correctly.

	Precision	Recall	F-score
source	0	0	0
cue	0	0	0
content	0.5	1.0	0.66
O	0	0	0

Table 3.5: Baseline performance for Test Set

The baseline system performance for token-wise labelling is expressed by,

$$\text{Accuracy} = 49.68\%$$

3.5.2 Sentence-wise Baseline

To check the sentence-wise labelling of an attribution relation, we prepared a sentence-wise baseline from the test set of data. For the sentence-wise baseline, we consider those sentences which have all tokens predicted as the actual label. Similar to the token-wise baseline, we consider all tokens in a sentence predicted as *content*. We observed that amongst 3022 sentences, 434 are found with all tokens labelled as *content*. The baseline system performance for sentence-wise labelling is expressed by,

$$\text{Accuracy} = 14.36\%$$

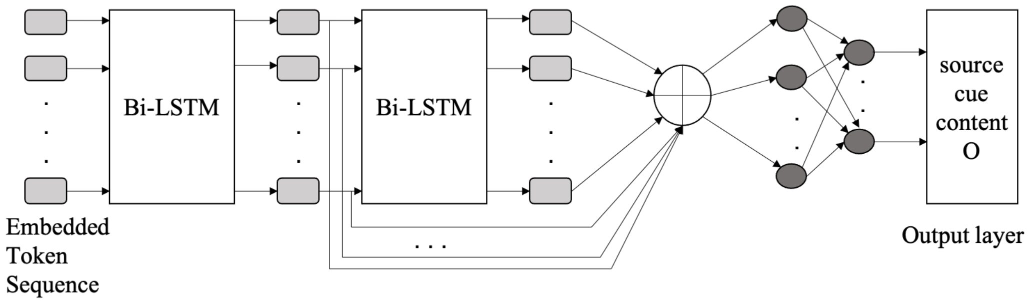


Figure 3.2: System architecture of AR model

3.6 System Architecture

In this section, I discuss the system architecture implemented to detect attribution relations in news articles. I build and implemented a new system for attribution relation detection using deep learning models that I made openly available.[2] I call this model for attribution detection- the "AR model" for short. I define the AR model as follows:

> Given, a sentence S from the body of a news article that consists of n tokens,
>
> $S = \{t_\text{i} \mid i \in [1, n]\}$ where, t refers to a token,
>
> an attribution relation detection model aims to tag each token in S where $tag \in \{source, cue, content, O\}$.

From the given definition, we are clear about the input and output of the AR model. Now, we discuss about the architecture of the AR model.

The input to the AR model is long texts that appeared as sentences in a news body. Here, I opt to use bi-directional long short-term memory (Bi-LSTM) network because of its capacity to handle long-term dependencies in sequence classification problems. The Bi-LSTM has shown good results in other NLP tasks. Although stacking layers can add more representational power (Hanselowski et al. 2018), multiple stacking may cause some representational degradation that can be resolved by having a residual connection (He et al. 2016). Thus, we implemented stacked Bi-LSTM with a residual connection.

[2] `https://github.com/NituB22/AR-model`

Word Embedding: As shown in the Figure 3.2, the input token sequence is represented in vectors using word embedding. I used a deep contextualized word representation called ELMO (Embedding from Language Models) by Peters et al. (2018). It is an embedding technique that can be considered as a function of an entire sentence containing that word. Thus, the same word can have different vectors in different contexts. For example,

Example 3.3.

i. After decades of backing mainstream politicians, European voters across the continent are increasingly empowering **right**-wing parties to upend Europe's long march toward a common economic, social and political union.

ii. Republican presidential candidate Donald Trump argued that fellow candidate Texas Senator Ted Cruz has "got to come a long way" on ethanol, "because he's **right** now for the oil" because "Oil pays him a lot of money.

iii. Flynn says Trump 'absolutely **right**' in saying generals reduced to rubble.

In Example 3.3, the word **right** in examples i, ii and iii have different meanings because it is used in different contexts. So, it is helpful if they are represented as three different vectors. The ELMO embedding makes it possible to consider context (words before and after **right**) to form the word vectors. In contrast, if we use an embedding like word2vec (Mikolov et al. 2013) that is context insensitive, embeddings for given word **right** in the given three examples are represented by the same vector.

As shown in the Figure 3.2, each sentence is represented as a sequence of tokens as $\mathrm{t}_1, t_2, \ldots, t_{50}$. The maximum length of a sentence or the sequence of tokens is 50. Each word is then represented by a 1024 dimensional vector using pre-trained ELMO embedding from TensorFlow Hub.[3]

[3] `https://tfhub.dev/google/elmo/2`

Neural network model: As shown in the Figure 3.2, each of those embedded token sequences is passed through two LSTMs in forward and reverse order. A Bi-LSTM works by processing a text sequence in both forward and reverse orders (Zhang et al. 2015). For instance, if there is a sentence "Jack said it is peaceful here" then LSTM with forward pass gets the sequence from "Jack" to "here". In contrast, the LSTM with reverse pass gets the sequence of words from "here" to "Jack". The importance of using bi-directional LSTM is to effectively capture the context of a word from both directions. The output vectors of forward and reverse LSTMs are concatenated and fed into another but, similar Bi-LSTM model. We implemented Bi-LSTM using Keras library.[4]

Similar to the first Bi-LSTM model, the second one also generates word vectors by concatenating the vectors from forward and reverse LSTMs. The final 1024 dimension word vectors from the first Bi-LSTMs, which is also known as residue, are directly added to 1024 dimension word vectors from second Bi-LSTMs. Using the residual network, the network is allowed to skip training of those layers that may not be useful and add no value in the overall accuracy. Concatenated word vectors are generated for 50 tokens ($\mathrm{t}_1, t_2, \ldots, t_{50}$) at different time steps $\mathrm{T}_1, T_2, \ldots, T_{50}$.

Sequence Tagging: As shown in Figure 3.2, the vectors for 50 tokens $t_1, t_2, \ldots, t_{50}$ from Bi-LSTMs are fed into a densely connected feed-forward neural network at different time steps $T_1, T_2, \ldots, T_{50}$. A softmax activation function is applied at the output layer for the final prediction. There are four output neurons representing the four tags/labels **source**, <u>cue</u>, *content* and O. Labels for each token in the sentence are predicted at different time steps.

[4] `https://keras.io/api/`

3.7 Results and Discussion

Recent works in automatic detection of AR in news articles by Pareti (2015) and Newell, Cowlishaw & Man (2018) implemented a pipeline system with three different classifiers to identify each AR component. The pipeline also consists of components that link the source, cue and content of each AR to automatically extract a complete AR. In contrast, I implement a system that can automatically identify all three AR components in a sentence as a single step. The use of deep learning models gives a simpler architecture for AR component detection. My system handles AR component detection as a sequence labelling task. In this section, I analyse how well the attribution detection works at both token and sentence levels. Additionally, I compare the model performance with the baseline previously discussed in Section 3.5. I am unable to compare my system with other systems (Newell, Cowlishaw & Man 2018) because of unavailability of their system. Additionally, so far to my knowledge there is no published works available that used the PolNeAR corpus for AR detection.

3.7.1 Token-wise Results

Here, we see how well the AR model performed for the token-wise labelling of news text. For the analysis, we use the test set of the PolNeAR corpus that is the unseen data for the model. The test set consists of 84 news documents. The confusion matrix represented by Table 3.6 shows the classification result of each token as any one of four given labels **source**, cue, *content* and O.

		Actual			
		source	cue	*conetent*	O
Predicted	**source**	4799	55	194	858
	cue	67	2967	477	1136
	content	129	160	32237	3401
	O	589	536	3013	21681

Table 3.6: Confusion matrix for Test set

Using the given confusion matrix we computed three different evaluation metrics to analyse the model performance to label each token correctly. The

AR model performance on the test set of the PolNeAR corpus is shown in Table 3.7.

	Precision	**Recall**	**F-score**
source	0.81	0.85	0.83
cue	0.63	0.79	0.70
content	0.89	0.89	0.89
O	0.83	0.80	0.81

Table 3.7: Token-wise AR model performance on Test set

The overall accuracy of the model for token labelling is as follows,

Accuracy = 85.32 %

3.7.2 Sentence-wise Results

Here, we evaluate how well the model performed at the sentence level. A sentence is said to be correctly predicted if all tokens in the sentence are correctly labelled as the actual tags. We see how well the AR model performed to correctly predict all tags of a sentence. Following are the results for sentence-wise prediction of token tags in the test set of the PolNeAR corpus.

Accuracy = 57.37%

3.7.3 Error Analysis for Sentence-level prediction

As the objective of this experiment is to extract an attribution relation, we mainly focus on correctly predicting the whole sentence. For error analysis, I manually analysed false cases in the validation set of the PolNeAR corpus. Here, “false cases” refer to those sentences which have at least one incorrect token prediction. We chose the validation set for error analysis such that if there will be any enhancement in the model, it will not introduce any bias to the model. This helps keep the unseen test set data safe and unbiased despite if any modifications have to be done to the model.

There are 1450 sentences with at least one mispredicted token tag. While analysing those data, we observed the following:

i. There are 462 sentences with all tokens labelled as *content* but, the model predicted all tokens in 45 sentences (that is around 10%) as O. While critically analysing, it is found that around ten percent of those mispredicted sentences are phrases with the number of tokens less than or equal to 4. For instance, *Thousands of shootings .*, *Probably not .*. This might be happening because there could be a large number of such phrases appearing in sentences with all token tagged as O.

ii. Around 4% of sentences with all tokens labelled as O are mispredicted with all tokens tagged as *content*. Amongst 32 mispredicted sentences, 10 of them are quotations that are sentences enclosed within inverted commas. For instance, "You walk down the street and you get shot ." While analysing those data, we observed that it could be due to a higher number of such quotations being tagged as *content*. 298 such quoted sentences are found in the gold data with all tokens tagged as *content*.

These observations show that similar texts are shared between *content* and O labels. If we could bind sentences with all tokens tagged as *content* to their respective attribution relations then this error could be handled. Here, I input a sentence to the system instead of an AR because there is a high variation in the lengths of attribution relations with the longest one with 1464 tokens and the shortest with 3 tokens in the training set of the PolNeAR corpus. In contrast, the longest sentence contains only 220 tokens. So, I opt to work with texts with low variation in their lengths. Additionally, my plan is to analyse how *content* component of ARs affects the stance detection problem. So, finding each AR separately is not mandatory for me as I can easily extract claims by taking tokens predicted as *content* by the AR model.

Note: In the examples, I represent attribution constituents as **source**, cue and *content*. The texts that are not part of an attribution relation are represented as O.

Amongst those 1450 mispredicted sentences, we randomly selected 300 sentences that are around 20% of data and analysed them manually. During the manual analysis, we observed the following:

i. We observed that 32 mispredicted cue labels are actually cues of nested attribution relations. The PolNeAR corpus doesn't have nested attribution annotations. In the case of nested ARs, they annotated **source** and cues appearing at the first hierarchy and rest texts as *content*. During error analysis, we observed two different scenarios of nested AR with mispredicted cue labels that are as follows:

   - Nested sources in the sentence. For example,

     Bloomberg reports *that the Republican nominee has either won or tied among the group of voters making $ 100,000 or more ,* according to **the Roper Center for Public Opinion Research** .

   - AR within another AR existing while considering the whole AR that is spanning over multiple sentences. Such cases are encountered when a cue is identified in sentences where all tokens are tagged as *content*. For example,

     “**WALLACE** : *Let 's start with breaking news on the debate story .After the Clinton camp* announced *that it was inviting billionaire and Trump critic Mark Cuban to sit in the front row at the debate , Trump invited Gennifer Flowers , who once had an affair with President Clinton , to also sit in the front row . And she has accepted that invitation . Two questions : Why would Trump do that ? And will Gennifer Flowers actually be there tomorrow night ?*

ii. 17 sentences are found with all tokens predicted as O but, actually seems to contain an attribution relation. The model predicts most of the attribution constituents of such sentences correctly. We can say that those ARs are the ones missed by annotators to annotate as ARs. It shows how good the AR model is because the model predicted ARs that are not identified by the annotators. For example,

   Annotated text: Figueroa said Stein complied with police orders and left the area .

Predicted text: **Figueroa** said *Stein complied with police orders and left the area* .

The following is an example with a pre-existing attribution relation but, the second AR was identified correctly by the AR model.

Annotated text: Many leaders of ethnic communities in the U.S. share these apprehensions , but **the Central and East European Coalition , which represents more than 20 million Americans** , wants *to preserve its non-endorsement policy* .

Predicted text: **Many leaders of ethnic communities in the U.S.** share *these apprehensions* , but **the Central and East European Coalition , which represents more than 20 million Americans** , wants *to preserve its non-endorsement policy* .

iii. Sentences containing long source spans which are usually longer than 20 tokens labelled as **source** have almost all labels mispredicted as O. For example,

The Democratic presidential nominee has been endorsed by **dozens of papers ranging from such expected backers as The New York Times to such once-certain Republican advocates such as The Dallas Morning News , the Arizona Republic and the Cincinnati Enquirer**.

3.8 Comparison with the Baseline

To evaluate how well the AR model is performing to label different constituents of attribution relations, we compare it with the baseline system that is explained in Section 3.5. The comparison between the baseline and the AR model performance is shown in Table 3.8. The data used for the comparison is the test set of the PolNeAR corpus.

	Baseline	**AR model**
Token-wise Accuracy	49.68%	85.31%
Sentence-wise Accuracy	14.36%	57.37%

Table 3.8: Comparison between the baseline and the AR model performance

Observing the given tables, we can say that the AR model outperformed the baseline system for both token-wise and sentence-wise predictions.

3.9 Broader Applicability of the AR model

In Section 3.8, I observed that the AR model shows promising results to label tokens of attribution relations. Here, I will analyse and validate whether the AR model is applicable to wider range of texts and beyond its data domain. This validation testing of the AR model will conclude whether the model is appropriate to use for other domain data. In the following section, we discuss the corpus from a different domain that will be used to validate that AR model, followed by a detailed analysis of the performance.

3.9.1 Data: Vaccination Corpus

I evaluate the performance of the AR model by testing it on the Vaccination Corpus[5] that has attribution annotations. The Vaccination corpus has 294 documents related to the vaccination debate collected from several sources including news, editorials, blogs, Wikipedia and a variety of health information dissemination websites (Morante et al. 2020). The corpus captured perspectives with three layers of annotated information that are attribution, claim and events. Amongst those, attribution is the layer of our concern.

I used the vaccination corpus as a test set to get the AR model performance that is trained using the PolNeAR corpus[6]. The PolNeAR corpus is explained in Section 3.3. The data distribution in the Vaccination corpus is shown in Table 3.9.

	Vaccination corpus
Documents	294
Total Attributions	4877
Total Sentences	23467
Sentences with attribution	6469

Table 3.9: Vaccination Corpus statistics

[5] `https://github.com/cltl/VaccinationCorpus`

[6] `https://github.com/networkdynamics/PolNeAR`

Table 3.9 shows that there are fewer numbers of attribution relations in the vaccination corpus despite the numbers of documents and sentences being higher in comparison to the PolNeAR corpus (See Table 3.3 for PolNeAR corpus statistics). It could possibly because the Vaccination corpus not only has news articles but, also other types of documents like blogs, editorials and so on. It shows that attribution relations appear more often in news documents. We found that 25 documents in the corpus have no attribution annotations. However, we included those documents for evaluation to analyse whether the AR model performance is affected by those documents. Additionally, we analysed the results excluding those data.

3.9.2 Annotation differences from PolNeAR corpus

We found several differences in the way attribution relations are annotated in the PolNeAR corpus and the Vaccination corpus that are as follows.

- The first difference is that, in cases where cues are realised by verbs, all pre- and post- modifiers, auxiliaries including modals, negative particles, adverbials and so on are excluded and only head verbs are annotated as cues in the Vaccination corpus. In contrast, the PolNeAR corpus has all such texts associated with verbs are annotated as cues. For example (with cues underlined),

  PolNeAR corpus: The boy was claiming it was a spider.

  Vaccination corpus: This country has reported cases of Zika virus infection in the past 9 months.

- The second difference is that, in the Vaccination corpus, punctuation markers such as semicolons and commas are annotated as cues only when no other lexical cue is available. In the PolNeAR corpus, both punctuation marks and lexical texts are annotated as cues even in presence of lexical cues. For example (with cues underlined),

  PolNeAR Corpus: His advice: “Run!”

  Vaccination Corpus: Dr. Ngare added: “The Catholic Church has been

> here in Kenya providing health care and vaccinating for 100 years for longer than Kenya has existed as a country."

Therefore, there are annotations in the Vaccination corpus which are annotated as O, although they would be annotated as cues in the AR model. This would lower the apparent precision of a model trained on the PolNeAR corpus when applied to the Vaccination corpus.

Analysis of cue distribution in PolNeAR and Vaccination corpora

In the PolNeAR and Vaccination corpora, we not only observed cue annotation differences but, also a difference in cue distribution that is affecting the cue prediction results. To analyse how cues distribution in those corpora might have affected the classification results, we extracted lists of the most frequent 30 cues from both PolNeAR and Vaccination corpora. The following are details of the top 30 occurring cues in PolNeAR and Vaccination corpora and cues are presented in the order of highest frequency to lowest. It means the first cue is the most frequent amongst all and the last one the least.

- PolNeAR corpus:

  say, tell, accord to, add, write, call, ask, show, note, :, suggest, argue, report, announce, have say, support, think, appear, seem, want, find, describe, in, cite, claim, see, believe, tweet, explain, acknowledge

- Vaccination corpus:

  say, recommend, know, think, tell, believe, accord to, ask, show, state, suggest, report, find, claim, decide, conclude, :, advise, admit, call, understand, declare, write, remember, blame, realize, discuss, wonder, explain, warn

The following cues are amongst most-frequent thirty cues that are common to both the Vaccination and PolNeAR corpus.

> say, think, tell, believe, accord to, ask, show, suggest, report, find, claim, :, call, write, explain

The following cues are in the Vaccination corpus top 30, but do not appear in the PolNeAR corpus:

> recommend, know, state, decide, conclude, advise, admit, understand, declare, remember, blame, realize, discuss, wonder, warn

Occurrences of given cues in the PolNeAR corpus is lower than that in the Vaccination corpus. 'recommend' is the second-highest occurring cue in the Vaccination corpus that appeared only 6 times in the PolNeAR corpus. Such cues degrade the AR model performance in the Vaccination corpus for cue detection because the AR model is trained with the PolNeAR corpus. The AR model may not have seen such cues more because of which it fails to identify them correctly in the Vaccination corpus.

3.9.3 Using the Vaccination corpus as the test set in the AR model

The AR model is trained with the training set of data from the PolNeAR corpus as described in Section 3.6. Details of the PolNeAR corpus, which is used to train and validate the AR model using its respective training and development sets, is discussed in Section 3.3.

The AR model is trained with 840 news articles related to Politics. The training set of PolNeAR corpus has around 30K sentences and around 20K attribution relations. The model is validated using a development set of 84 news articles from the PolNeAR corpus that has more than 3K sentences and 2K attribution relations. Thereafter, the AR model performance is tested on a completely different corpus- the Vaccination corpus.

3.9.4 Token-wise Results

Table 3.10 shows how well the AR model classified each token of the Vaccination corpus that is represented in the format of a confusion matrix.

The performance of the AR model in the Vaccination corpus is expressed

		Actual			
		source	cue	*content*	O
Predicted	**source**	7136	14	2280	6519
	cue	53	2678	1899	9738
	content	1369	886	75883	114728
	O	1404	263	29870	253045

Table 3.10: Confusion matrix of Vaccination corpus data prediction by AR model

using three different metrics- precision, recall and F-score that are given in the Table 3.11. The table shows the model's performance for token labelling.

	Precision	**Recall**	**F-score**
source	0.45	0.72	0.55
cue	0.19	0.70	0.29
content	0.39	0.69	0.50
O	0.89	0.66	0.76

Table 3.11: Token-wise AR model performance on Vaccination corpus

The AR model performance for cue labelling is not as good as other label classification that is anticipated previously during the data analysis discussed in Section 3.9.2. It is possibly due to different domain information and writing style used in the Vaccination corpus texts.

Excluding 25 No AR documents There are 25 documents in the Vaccination corpus that have no attribution relations. Those 25 documents have 433 sentences and around 8.5K tokens. The confusion matrix in Table 3.12 shows the AR model performance on the Vaccination corpus while excluding data without any attribution relations.

		Actual			
		source	cue	*content*	O
Predicted	**source**	7136	14	2280	6369
	cue	53	2678	1899	9616
	content	1369	886	75883	113084
	O	1404	263	29870	246408

Table 3.12: Confusion matrix of Vaccination corpus data excluding 25 Documents with No ARs

The token-wise labelling results in Table 3.13 shows the performance of the AR model in the Vaccination corpus while excluding documents that have no attribution relations.

	Precision	**Recall**	**F-score**
source	0.45	0.72	0.55
cue	0.19	0.70	0.30
content	0.40	0.69	0.50
O	0.89	0.66	0.75

Table 3.13: Token-wise AR model performance on Vaccination corpus excluding 25 Documents with No ARs

According to Table 3.11 and Table 3.13, we can say that there is no significant improvement on the *cue* prediction. However, the performance of *O* labelling decreased slightly because exclusion of all *O* documents should decrease false-positive cases for all other labels hence, improving their results.

3.9.5 Sentence-wise Results

The following are the sentence-wise prediction results of the AR model on the Vaccination corpus:

Number of sentences with all correct token predictions = 10250

Number of sentences with all O = 7670

Sentence-wise accuracy = 43.67%

Excluding 25 No AR documents While excluding 25 documents from the Vaccination corpus that have no attribution relations, the following sentence-wise results are obtained:

Total sentences = 23033

Number of sentences with all correct token predictions = 9992

Number of sentences with all O = 7416

Sentence-wise accuracy = 43.38%

In the case of 25 documents with no ARs, sentence-wise results show that 254 out of 433 sentences are correctly identified as all *O* sentences by the model. It means only 59% of sentences from excluded documents were labelled correctly by the AR model. It implies that the AR model is not able to filter all documents that have no attribution relations.

3.9.6 Comparison with the Baseline and Test set results of PolNeAR corpus

As the baseline system, we considered the highest appearing label as the predicted label for all tokens. As *content* is the highest appearing token amongst AR constituents, we label all tokens as *content* and then analyse the token-wise and sentence-wise accuracy. There are around 508K tokens in the vaccination corpus amongst which around 110K are labelled as *content*. There are around 2.1K sentences with all tokens labelled as *content*. Table 3.14 shows the comparison of token-wise and sentence-wise accuracy of the Vaccination corpus results with its majority class baseline; and the test set results of PolNeAR corpus.

	Vaccination	**Baseline (Vaccination)**	**Test Set (PolNeAR)**	**Baseline (PolNeAR)**
Token-wise Accuracy	66.71%	21.65%	85.16%	49.68%
Sentence-wise Accuracy	43.67%	9.20%	56%	14.36%

Table 3.14: Comparison of Vaccination corpus results with its majority class baseline and Test set results of PolNeAR corpus

Table 3.14 illustrates that the AR model performed is promising in the Vaccination corpus while comparing with the baseline system. However, the results are not up to the level of the PolNeAR corpus test set. It could be due to the difference in the domain of data present in the Vaccination corpus. Nevertheless, the AR model performance is encouraging in a completely different domain dataset. Additionally, poor baseline results imply that the corpus has a small amount of attribution data in comparison to the overall data in the corpus. It could also be a reason for the poor performance of the

AR model in the Vaccination corpus.

3.9.7 Discussion

The results of the AR model on the Vaccination corpus are promising for the broader applicability of the AR model. However, the results are not as good as those in the test set of the PolNeAR corpus. This could be due to the much greater range of topics covered in the Vaccination corpus. Additionally, the PolNeAR corpus has news articles that were published in different news media whereas the Vaccination corpus contains many different genres of documents, including news articles, blog posts, Wikipedia texts, editorials and so on. The distribution of annotation tags in the two corpora also suggests that news articles contain more attribution relations than the other text types in the Vaccination corpus.

The sentence-wise tagging performance is encouraging despite the presence of substantial unrelated data in the Vaccination corpus. The data shows that around 65% of data in the PolNeAR corpus are ARs, whereas the Vaccination corpus has around 25% of data as ARs. The non-AR data may lead to a lot of false predictions. Similarly, the model performance for cues token tagging is the worst case as predicted, this is at least partly due to annotation differences between the PolNeAR and Vaccination corpora.

I excluded 25 documents with no attribution relations and analysed the results. I observed no significant improvement in the AR detection although I observed a slight decrease in false predictions of AR constituents as O. I observed around 80% of tokens and 60% of sentences in those 25 documents are predicted correctly by the AR model. This further implies that the model can satisfactorily filter documents with no attribution relations.

In conclusion, despite of being tested in a different domain data like the Vaccination corpus, the AR model showed an encouraging performance for its broader applicability in other corpora.

3.10 Conclusion

In this chapter, we build and implement a new model for attribution relation detection that showed promising results for correctly classifying three different components of an attribution relation. To evaluate the performance of the AR model, we compared it with a baseline system and we observed that the AR model results are promising. To analyse the broader applicability of the AR model, we used it to detect ARs in a different corpus called the Vaccination corpus that not only has news documents but, also several types of texts related to vaccination. The broader applicability of the AR model is validated by its encouraging results in the Vaccination corpus. The purpose of this broader applicability testing is to ensure the effectiveness of the AR model beyond the data domain in which it is trained. Furthermore, our plan is to use the AR model to detect attribution relations in a corpus that has stance annotation but, doesn't have attribution annotations. By detecting ARs in a document, we can get sources tagged as **source** and their respective claims which are tagged as *content* that are further used in the stance detection task to assess their roles.

Through this chapter, I present my system for attribution detection that can tag components of attribution relations in news texts. For the AR model's broader applicability, I also evaluated the model's usability in a different domain data.

Chapter 4

Detecting Stance using Attribution Relations

4.1 Introduction

The headline can not always be a reflection of its associated news body. We can identify and flag such news articles with misleading headlines by finding the stance of the news body relative to the headline. Küçük & Can (2020) defined stance detection as a classification task where the stance of the body towards the claim of the headline is in the form of a category label. The category label can be agree, disagree, discuss or unrelated that are explained previously in Section 2.1.

The stance information, where a news body is disagreeing with the headline, can be useful to indicate possible misinformation to readers. So far to my knowledge, except for the work by Sepúlveda-Torres et al. (2021), all works using the FNC-1 corpus for stance detection used the whole news text for the classification task. However, several works (Ghanem et al. 2018, Conforti et al. 2018) have highlighted the computational complexity which arises from the length asymmetry between the news headline and the associated body. Sepúlveda-Torres et al. (2021) proposed to mitigate the length asymmetry by representing the news body with its summary that is extracted using an off-the-shelf tool. My approach of representing a news body by attribution relations is a stepping stone to the solution. Attribution relations containing important pieces of information from the news body could be useful to detect the stance. Thus, in this chapter, I am working on the following research question:

> RQ2: Are attribution relations useful to detect the stance of a news body to its headline?

So far, to my knowledge, this work is the first one to study the usefulness of attribution relations for detecting the stance of news bodies towards their headlines.

Headline	Tropical spider burrowed under man's skin through appendix scar and lived there for THREE DAYS
Article	The 21-year-old was on his first trip to Bali. He told News Corp yesterday *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin*. There's just one problem. *Spiders*, according to Perth arachnid expert Dr Volker Framenau, *don't burrow in skin*. *"They don't have the tools, the armature, to do this sort of stuff,"* Dr Framenau said.
Stance	disagree

Table 4.1: Example with headline-article pair dissonance, and *contents* as *claims*

For example, consider the news article shown in Table 4.1, taken from the Fake News Challenge (FNC-1) corpus (Pomerleau & Rao 2017). In the example, a 21-year-old boy claims that doctors pulled a tropical spider from his skin. On the other hand, Perth arachnid expert Dr Volker Framenau claims that spiders can't burrow in skin because "they don't have the tools. . . to do [such] stuff". In the corpus, the stance of the given example is disagree. However, in this case, two different sources are making conflicting claims, one of which agrees with the headline, and one of which disagrees. Generally, news reporting is based on information provided by different sources. Therefore, it is important to critically deal with such claims from sources rather than using the entire news article text for stance classification. We can get such sources and their respective claims by detecting attribution relations in the news body.

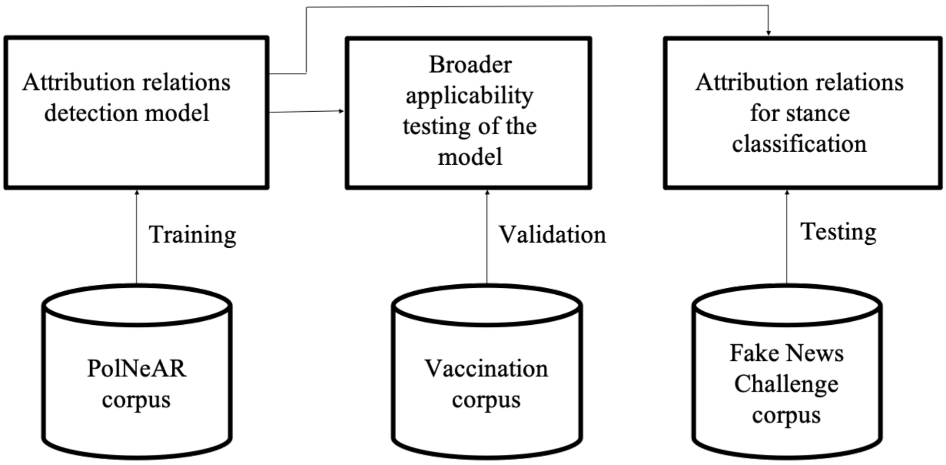

Attribution relations detection model
Broader applicability testing of the model
Attribution relations for stance classification
Training
Validation
Testing
PolNeAR corpus
Vaccination corpus
Fake News Challenge corpus

Figure 4.1: A pipeline for attribution detection in FNC-1 corpus with associated corpora

To the best of my knowledge, there is no corpus available that contains both attribution relation and stance annotations in news articles. I, therefore, implement a model for attribution detection using the PolNeAR corpus that has attribution annotations; which is previously described in Chapter 3. Figure 4.1 shows the task flow along with the corpora used at different stages. The purpose of the pipeline in the Figure 4.1 is to detect ARs in the FNC-1 corpus that has only stance annotations.

In this chapter, I focus on using attribution relations detected in the FNC-1 corpus for the stance detection task. A detailed literature review on the stance detection is given in Chapter 2.

4.2 Data

For stance classification, we use the FNC-1 corpus (Pomerleau & Rao 2017) introduced during a Fake News Challenge that used the same 300 claims and 25K articles used in the Emergent corpus (Ferreira & Vlachos 2016). There are four headline-article stance labels used in the FNC-1 corpus: agree, disagree and discuss, and unrelated for cases where the body text discusses a different topic from the headline.

The FNC-1 dataset consists of 75,385 samples amongst which 54,894 are annotated as unrelated and 20,491 are annotated as agree, disagree or discuss. Table 4.2 shows the percentage distribution of each data class in the corpus.

	Considering	
Stance	**All data (%)**	**Related data (%)**
agree	7.4	27.2
disagree	2.0	7.5
discuss	17.7	65.2
unrelated	72.8	-

Table 4.2: Stance label distribution in FNC-1 corpus considering all data; and only related data

There are respectively 49,972 and 25,413 news articles in the training and test set of the FNC-1 corpus. The training set is prepared with 1,648 unique headlines and 1648 unique articles combined to form more than 49K headline-article pairs. The stance label distribution in the training set is shown in Table 4.3. Similarly, the test set of the FNC-1 corpus is made with 904 unique articles that do not match with that in the training set. The stance label distribution in the test set is given in Table 4.4.

Stance	**headline-article**	**%**
agree	3,678	7.3
disagree	840	1.6
discuss	8,909	17.8
unrelated	36,545	73.1

Table 4.3: Stance label distribution in the Training set of FNC-1 corpus

Stance	**headline-article**	**%**
agree	1,903	7.4
disagree	697	2.7
discuss	4,464	17.5
unrelated	18,349	72.2

Table 4.4: Stance label distribution in the Test set of FNC-1 corpus

From Table 4.2, we can know that the FNC-1 corpus is highly biased towards the stance label unrelated that will certainly affect the performance of a stance detection model. It is because unlike the unrelated class, other stance classes have very less representative samples. Therefore, I decided to opt for a two-stage stance detection system. The first stage deals with filtering related data. At the second stage, correctly classified related data from the first stage is further classified to any one of given three labels agree, disagree or discuss. Detailed literature for implementing a two-phase architecture is given in Chapter 2.

4.3 Input and Expected Output

Let us consider the example given in Table 4.1. The inputs to the stance detection model are the headline and attribution relations detected in the news body using the AR model. Figure 4.2 shows the input and the output of my stance detection model. An attribution relation in the input is represented by its three components **source**, cue and *content*. I refer to an attributed content as a *claim*. The output of the system is a stance label that could be agree, disagree, discuss or unrelated. In the given example, the stance label is disagree. Furthermore, besides using ARs for the stance detection, we also analyse how the system performs while only using *claims* in the input.

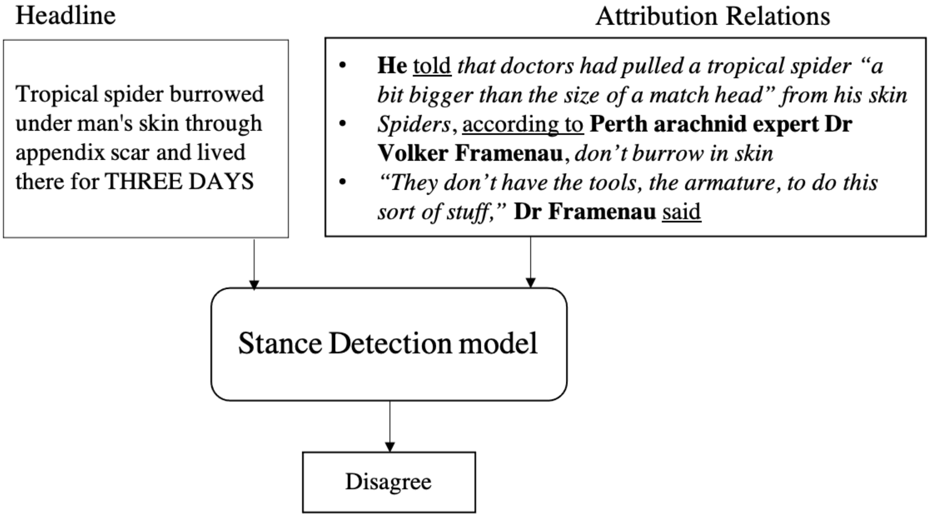

Headline
Tropical spider burrowed under man's skin through appendix scar and lived there for THREE DAYS
Attribution Relations
He told *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin*
Spiders, according to **Perth arachnid expert Dr Volker Framenau**, *don't burrow in skin*
"They don't have the tools, the armature, to do this sort of stuff," **Dr Framenau** said
Stance Detection model
Disagree

Figure 4.2: Input and Output of the Stance Detection model

4.4 Baseline

The Fake News Challenge[1] provided a baseline system. The baseline implemented a Gradient Boosting classifier with features like n-gram co-occurrence between the titles and articles using both character and word n-grams. The feature set also include other hand-crafted features such as the existence of highly polarized (such as fake, hoax) and refutation words. Weighted scoring is done by giving 25% weighting to classify between unrelated and *related* classes. If the label is *related*, the data is further processed for the final stance labelling as agree, disagree or discuss with a weighting of 75% for the correct predictions. Pomerleau & Rao (2017)'s baseline system has an accuracy of about 75.2% and a macro F1 score of 46.9%. Additionally, I consider top three winning systems from the fake news challenge, discussed earlier in Section 2.5 as baseline systems for my work.

4.5 System Architecture

I implemented a two-stage approach such that unrelated labels are filtered at the first stage and only related data are handled at the second stage as shown in Figure 4.3.

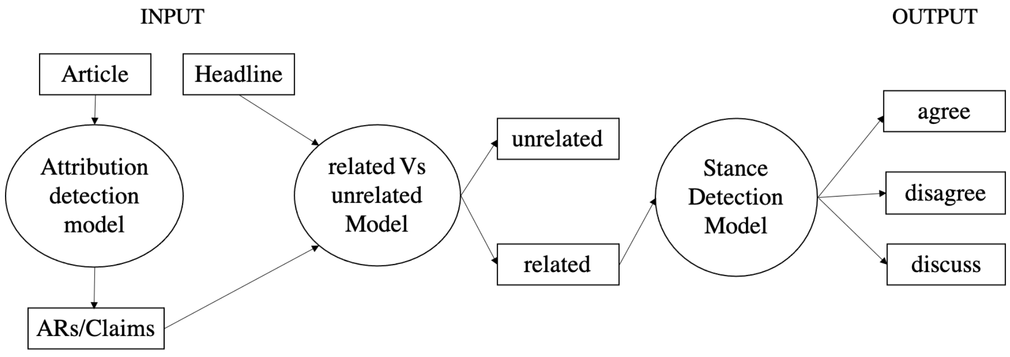


Figure 4.3: A two-stage Stance Detection architecture

As we can see in Figure 4.3, first I implemented the attribution detection model to the FNC-1 corpus. The AR model labels each token in news bodies of FNC-1 corpus as one of four labels: **source**, cue, *content* or O. The label

[1] http://www.fakenewschallenge.org/

O represents that the token does not belong to an attribution relation. The other three labels represent respective components of an attribution relation. There are respectively 1648 and 904 unique news articles in the training and test sets of the FNC-1 corpus; details of which are provided in Section 4.2. Therefore, I find token-wise labels in those unique articles that are later combined with several headlines to form a news article in the FNC-1 corpus. The implementation of the attribution detection model on the FNC-1 corpus resulted in the labelling of each token as reported in Table 4.5.

	Training set	**Test set**
source	45,810	21,800
cue	28,304	14,746
content	329,943	178,125
O	314,165	175,894
Total Sentences	33,729	18,190
Sentences with AR	23,053	12,481

Table 4.5: AR components detected by the AR model in FNC-1 corpus

In Table 4.5, we can see that around 69% of total sentences have attribution relations that are comparable to the PolNeAR corpus which has 71% of sentences with ARs. Furthermore, the most frequent token in the FNC-1 corpus is *content* as in the PolNeAR corpus. I anticipated such results because FNC-1 has news articles where attribution relations appear more in comparison to other types of articles, as observed in Section 3.9.1.

Once all the attribution relations are detected in the FNC-1 corpus, we implement a binary classifier to filter unrelated class articles. In the first phase, there are only two classes, one with unrelated labelled data; and another with agree, disagree and discuss labelled data. Slovikovskaya & Attardi (2020) argued that the stance detection task can benefit from transfer learning in pre-trained transformers. Therefore, I used Simple Transformers by Rajapakse (2017) that is built on top of Hugging Face transformers to implement pre-trained models.[2] The input to my stance classification model is a pair of headline and concatenated attribution relations extracted by the AR model from FNC-1 news bodies. The output of the model is one

[2] https://github.com/huggingface/transformers

of the two stance labels unrelated or *related*. At both the first and second stages, I implemented a large model of Robustly Optimised BERT approach (RoBERTa) (Liu et al. 2019) using the Simple Tranformer Library for the stance detection task.[3] I trained transformers for 3 epochs with a batch size of 4 and a learning rate of $1e^{-5}$. At the second stage, the data classified as related from the first model are used as input to the stance detection model. The output of the stance detection model is any one of three given labels agree, disagree or discuss. At the second stage, I implemented the same deep learning model which was used at the first stage.

As shown in Figure 4.3, the model has two stages that let me evaluate the stance detection model (the second stage) in at least two ways:

i. in the context of the output of the related vs unrelated detection

ii. in isolation (i.e. with the assumption that the first step produces correct output)

I opt for (i) because it does not exclude unrelated class from consideration making it possible to compare my system performance with current best systems. So far to my knowledge, there is only a work by (Conforti et al. 2018) that evaluated their stance detection system without considering the unrelated class.

4.6 Results and Discussion

This section discusses the results and evaluation of my two-stage stance detection system. The unseen test set of the FNC-1 corpus is used to evaluate the performance of the system. The section ends with a comparison of my stance detection system results with baseline and state-of-the-art systems.

4.6.1 Using only content in input

Firstly, I used headlines and concatenated *content* for each news article as the input to the two-stage stance detection system. I consider *content*

[3] https://simpletransformers.ai/

components of an attribution as potential claims from different sources in that AR. The overall labelling done by the two-stage system is represented as a confusion matrix shown in Table 4.6.

		Predicted			
		agree	disagree	discuss	unrelated
Actual	agree	1210	51	537	58
	disagree	148	296	198	27
	discuss	488	119	3638	128
	unrelated	59	9	166	17819

Table 4.6: Confusion matrix while using only *content* in the input

The two-stage stance detection model performance while using only *content* in the input is given in Table 4.7.

	Precision	**Recall**	**F-score**
agree	0.64	0.65	0.64
disagree	0.62	0.44	0.52
discuss	0.80	0.83	0.82
unrelated	0.99	0.99	0.99

Table 4.7: A two-stage stance detection model performance using only *content* in the input

4.6.2 Using Attribution Relations in input

Secondly, considering that all components of an attribution relation are equally important to convey meaningful information to the reader; I used attribution relations instead of claims in the stance detection system. Here, the headline and concatenated attribution relations for each news article are fed to the two-stage stance detection system. The confusion matrix of stance detection while using attributions is given in Table 4.8. The confusion matrix shows that the model is confident at predicting frequently appearing unrelated class data. However, the least appearing stance classes like agree and disagree are frequently mis-predicted as each other or as the discuss class.

The two-stage stance detection model performance while using attribution relations as input is given in Table 4.9. The table shows that my stance

		Predicted			
		agree	disagree	discuss	unrelated
Actual	agree	1254	67	491	46
	disagree	150	311	195	13
	discuss	331	103	3878	64
	unrelated	58	5	124	17861

Table 4.8: Confusion matrix while using Attribution Relations as input

detection model is almost 100% efficient at predicting the highly occurring unrelated class data. The model performance degrades in minority classes (disagree and agree) due to the presence of less samples in comparison to other classes.

	Precision	**Recall**	**F-score**
agree	0.70	0.67	0.69
disagree	0.64	0.46	0.54
discuss	0.83	0.89	0.86
unrelated	0.99	0.99	0.99

Table 4.9: A two-stage stance detection model performance using Attribution Relations as input

4.6.3 Comparison with other best systems

Table 4.10 shows a comparison between my experimental results and other best systems for stance classification in the FNC-1 corpus, including the baseline. The stance detection system by Sepúlveda-Torres et al. (2021) utilised summary instead of whole news texts in the input for the stance classification. I used a similar method however, I used attribution relations in news bodies instead of their summaries to detect the stance.

My stance detection model outperformed systems by Hanselowski et al. (2018) and Roy et al. (2022) including all baseline systems. Hanselowski et al. (2018) used stacked Bi-LSTMs and used whole news body in their input. Roy et al. (2022) implemented a three-stage approach for stance classification. Better performance of my system implies that attribution relations in news bodies are important bit of information that can represent the news bodies. My two-stage model outperformed the three-stage model

	class-wise F-score				
	agree	**disagree**	**discuss**	**unrelated**	**macro F**
Baseline	0.17	0.01	0.72	0.97	0.46
Yuxi Pan (2017) (1st FNC)	0.53	0.03	0.76	0.99	0.58
Andreas Hanselowski (2017) (2nd FNC)	0.48	0.15	0.78	0.99	0.60
Benjamin Riedel (2017) (3rd FNC)	0.47	0.11	0.74	0.98	0.58
Hanselowski et al. (2018)	0.50	0.18	0.75	0.99	0.60
Zhang et al. (2019)	0.67	**0.81**	0.83	0.99	**0.83**
Slovikovskaya & Attardi (2020)	0.70	0.58	0.84	0.99	0.78
Sepúlveda-Torres et al. (2021)	**0.74**	0.64	**0.86**	0.99	0.80
Roy et al. (2022)	0.53	0.23	0.75	0.97	0.62
Using only *content* for Stance	0.64	0.52	0.82	0.99	0.74
Using attributions for Stance	0.69	0.54	**0.86**	0.99	0.77

Table 4.10: Comparing my system with the baseline and current best systems

by Roy et al. (2022) that also shows using attribution relations is effective. My model does not have a performance like that of Zhang et al. (2019) and Sepúlveda-Torres et al. (2021). It might be because of their respective use of a two-layered neural network architecture with a regularisation technique and use of summary extracted using a powerful off-the-shelf tool along with word features that are not used in my system. Regarding my work, using attribution relations slightly outperform against a *content* only solution. It might be because remaining two components of ARs which are **source** and cue contain shorter texts compared to a *content*. Furthermore, *content*s consist of claims made by different sources, containing information directly related to the event or incident discussed in the news article. The result of my model while using only *content*s out-performed works by Hanselowski et al. (2018) and Roy et al. (2022). It shows that *content* are self-sufficient to effectively detect the stance in a news article. It also implies that amongst AR components, *content* carries the most useful information for the stance detection. Although *content* shows high potential to be useful in the stance detection, remaining AR components like **source** and cue are equally likely to be highly effective if we can consider their features than just using their contents. The effectiveness of **source** expertise for the stance detection is assessed in Chapter 6.

Practically, all stance classes agree, disagree and discuss are important to be tackled carefully. The unrelated class is not realistic as it contains news articles that have topically irrelevant headline-body pairs. The disagree class is important to be identified to flag news articles such that any reader gives more attention to read the news article thoroughly. Such system that only focuses on the disagree class might be useful to flag news articles with possible misinformation. However, such a system might not be effective in other application areas of the stance detection. For example, a stance detection model with a good performance only for disagree class might not be useful for a news editor to find an appropriate headline for a news article where a headline should exactly reflects its associated news body. Furthermore, an effective model only for the disagree class might not be appropriate to be used to assess public opinions on some topic like Brexit. Thus, all stance classes are equally important to be identified correctly by the model for its wide usage. From Table 4.10, we can see that the current best system by Zhang et al. (2019) can be the best to tackle misinformation with the highest performance for the disagree class. However, Zhang et al. (2019) can not be the best for other domains such as in public opinion assessment. My stance detection model outperformed the work by Zhang et al. (2019) for agree and discuss classes with the same performance for the unrelated class. It implies that although my model has a poor performance than the state-of-the-art for the disagree class, it can have a broader applicability because of its better performance on other remaining stance classes.

My results reported in Table 4.10 shows a promising contribution of attribution relations in the stance detection task. My results outperformed all baselines that include the majority class baseline and the top three winning results of the fake news challenge (Pomerleau & Rao 2017). My results are encouraging. Despite of having an improvement space for agree and disagree classes, my model is comparable to the best systems for discuss and unrelated classes. It illustrates that representing a news article by attributions not only reduces the headline-article asymmetry but, is also useful to classify the stance. It further implies that attribution relations contain key information of a news body.

4.7 Error Analysis

In this section, I discuss about misclassified results of agree, disagree and discuss labelled data. I omitted unrelated class from this analysis because the first stage model is almost 100% efficient. In this error analysis, our major focus is on disagree class because its results are poor in comparison to other stance classes. During error analysis of each stance class results, I observe the following.

4.7.1 disagree **class**

There are low false positive cases in the disagree class. In Table 4.6 and Table 4.8 we can see that a very low number of unrelated, agree and discuss cases are mis-classified as disagree class. This implies that data in the unrelated, agree and discuss classes are not similar to that in disagree class.

Additionally, there are high false negatives cases in the disagree class. As per Table 4.6 and Table 4.8, we can say that many disagree stance labelled data are mis-classified as agree and discuss classes, resulting in high number of false negative cases. This implies that disagree labelled data share features with agree and discuss classes. Therefore, in Table 4.7 and Table 4.9, we can see that recall for disagree class is lower than the precision. The system failed to correctly identify many of disagree labelled news articles.

I evaluated 90 such news articles which is 26% of such data that are misclassified to agree or discuss classes. I observed the following:

Negation in the headline

49 news bodies are associated with headlines that contain negation words like *no, not*. This is the case in around 55% of the total news articles I evaluated. For example,

Example 4.1.
Headline: *Sorry, Argentina's President Didn't Actually Adopt a Jewish Werewolf*

Computationally detected AR extracts from News body:
The President of Argentina , Fernandez de Kirchner , has Jewish godson to prevent him from becoming a werewolf
According to the legend the seventh son of a family will transform into 'El Lobison', a werewolf like creature , on the first Friday after the boy's 13th Birthday , and will continue to turn into a blood-thirsty , baby eating werewolf
The President has said that Yair is the first Jewish boy to take part in the ceremony

The news article of Example 4.1 is given in Appendix A.1. Here, the news body associated with this headline has texts like *The President of Argentina , Fernandez de Kirchner , has Jewish godson to prevent him from becoming a werewolf*. This shows that despite there being a clear disagreement between the headline and the body, the stance is classified as agree.

One of the reasons for such results can be the much lower number of disagree labelled data in the training set, in comparison to the rest of the labels as shown in Table 4.3. Our model might not have got enough cases to detect such disagreements.

Fake cues in the headline

In 16% of data (14 news articles), headlines contain cues like *fake, hoax, false* that act as an opposing anchor concerning news body. Half of such headlines have negating cues like *no, not*. For example,

Example 4.2.
Headline: *Report: Woman who claimed to have the third breast added is fake*

Computationally detected AR extracts from News body:
A Florida massage therapist said she paid $ 20,000 for a third breast in hopes of becoming less attractive to men
" I don't want to date anymore ", Jasmine Tridevil told

She said she contacted more than 50 doctors before she found a surgeon willing to perform the operation

The news article of Example 4.2 is given in Appendix A.2. The associated news body contains claims by a woman about the surgery, why and how she did that.

The reason behind this misclassification can be the association of the same headline with several other news bodies that are labelled as agree or discuss in the FNC-1 corpus. As shown in Table 4.3, only 2% of data in the FNC-1 corpus belongs to disagree class that makes the training of the model not effective for the disagree class.

Preventive and Opposing words in the headline

There are other different cues like preventive words (for instance: *stop, prevent, ban*) and opposing words (for instance: *defend, denies*) that act as an anchor of disagreement in the headline and thus, creates dissonance with its associated body. For example,

Example 4.3.

Headline: *Saudi Airlines to ban gender-mixing seating*

Computationally detected AR extracts from News body:
According to an airline source quoted in stories the opposite Saudia Airlines the state - run airline of Saudi Arabia does not have plans to separate passengers based on gender

The news article of Example 4.3 is given in Appendix A.3. In the given example, the news body contains claims of having no plans to separate passengers based on gender. It makes the news body disagree with the headline. Despite this, such texts are easy to spot with disagreement, they are not classified as disagree. This shows that preventive and opposing word features could be helpful to make disagree class specific predictions.

Example 4.4.

> Headline: *Jasmine Tridevil: Woman with three breasts denies surgery hoax claims*
>
> Computationally detected AR extracts from News body:
> *she revealed she paid thousands of dollars to get a third breast surgically attached to her chest*
> *Florida woman Jasmine Tridevil claims that the surgery is a fake and she made it all up*
> *Surgeons have also dismissed the possibility of it being real*
> *New York plastic surgeon Matthew Schulman told: "[I] believe 100 per cent that this is a hoax that everyone is falling for,"*

The news article of Example 4.4 is given in Appendix A.4. Here, the news body mentions the claims about the woman's surgery and surgeons claiming that third-breast implantation is a hoax story. The woman also claimed that the surgery story is fake. These all information are disagreeing with the claim in the headline. Amongst my evaluation samples, there are 6 news articles with preventive words and 2 news articles with opposing words in their headlines.

1-AR news bodies

There is some interesting patterns where news articles with one or two sentences are mis-classified. Despite being very easy for humans to find the disagreement between the headline and the body, my model didn't classify such news articles correctly. For example,

> **Example 4.5.**
> Headline: *Report: White House Chief Of Staff Denis McDonough: No Threats Were Made To Foley, Sotloff Families Over Possible Ransom*
>
> Computationally detected AR extracts from News body:
> *The US threatened to prosecute James Foley 's family over ransom payments*

The news article of Example 4.5 is given in Appendix A.5. Here, the news body clearly disagrees with the headline. While analysing the training set, I observed that there are 620 news articles that have only an attribution relation in its body. Amongst those news articles, only 28 articles belong to the disagree class. That is less than 5% of the data. These numbers illustrate the presence of very few training samples to correctly classify the new data.

Headlines with no claims

Amongst 90 news articles, I observed that 2 articles have headlines that contain no claims. For instance,

> **Example 4.6.**
> Headline: *Giant Crab*
>
> Computationally detected AR extracts from News body:
> *Quinton Winter told that he'd spotted a giant crab in the mouth of Kent harbor while on vacation with his son last year*
> *"It had glazed blank eyes on stalks, swiveling wildly and it clearly was a massive crab with crushing claws," he said*
> *"The idea of a giant 'crabzilla' would [be] very exciting. Unfortunately, I think this is a hoax," Dr. Verity Nye, Ocean and Earth Science researcher at Southampton University, told "I don't know what the currents are like around that harbor or what sort of shapes they might produce in the sand, but I think it's more conceivable that someone is playing about with the photo."*

The news article of Example 4.6 is given in Appendix A.6. In the given example, a man claims that he spotted a giant crab at a harbour. In contrast, an Ocean and Earth Science expert claims that the story is a hoax and the crab related images are edited.

It was difficult for me to decide what the news body should contain to disagree with such headline. There is no overview paper or any annotation guidelines for the FNC-1 corpus available so far that is also mentioned by Hanselowski et al. (2018) in their work of stance detection.

4.7.2 agree and discuss classes

News articles in the agree and discuss classes of the FNC-1 corpus are indistinct. According to results in Table 4.6 and Table 4.8, we can say that agree and discuss classes share many features as there are lots of data misclassified in between those two classes. While manually analysing agree and discuss class news articles, it is difficult even for humans to distinguish between them because no annotation guidelines are provided by Pomerleau & Rao (2017). According to Pomerleau & Rao (2017), in the agree class, the body text agrees with the headline and in the discuss class the body text discusses the same topic as the headline, but does not take a position. An example with the same headline and different body texts that are misclassified by the model is given in Table 4.11.

4.7.3 Discussion

From the error analysis, we can say that the highly unbalanced FNC-1 corpus made it difficult for the machine learning model to correctly classify disagree labelled data. Additionally, I also observe that specific features like negation and fake cues in the headline are not handled properly in my system. One possible solution to this problem can be using task-specific features. For example, disagree and agree labelled news articles should contain texts that should show a visible stance to its headline. The three-stage pipeline model proposed by Roy et al. (2022) can be useful because we can implement stance specific features at the third stage after filtering unrelated and discuss classes at the first and second stages respectively.

4.8 Conclusion

In this chapter, the results illustrate that attribution relations in a news body are useful to classify the stance of the news body to its headline. The *content* components of an AR while used alone, that I also referred to as claims, also show promising results in the stance detection. It shows that attribution

Headline	Argentina’s President Adopts Young Boy so He Won’t Turn Into Werewolf	
Stance	agree	discuss
News Body	According to legend the seventh son born to a family turns into a ferocious " el lobison " or werewolf on the first Friday after his 13th birthday the fear of the people in the 19th century who believed their sons could turn into werewolfs formally to adopt daughters Tawil’s parents wrote to the president in 1993 for their son to be the first Jewish boy to be adopted and they got their wish this year according to the Jewish Telegraphic Agency Fernandez de Kirchner tweeted : " I didn’t know it but his visit coincided with the Hanukkah celebration . The father said it wasn’t a coincidence . " She added the meeting with him and his family was a " magical moment "	This is actually a thing that happened. According to The Independent Argentina’s president, President Christina Fernandez de Kirchner adopted a boy named Yair Tawil as her godson so that he would not turn into a werewolf According to Argentinian folklore the seventh son born to a family turns into the feared " el lobison " the first Friday after boy’s 13th birthday the legend says a at during every full moon , doomed to hunt and kill before returning to human form the lobison was said to be unnaturally strong and able to spread its curse with a bite Because some people actually believed this hundreds of years ago , they started killing babies. presidents starting the seventh born boys of families. According to The Jewish Telegraphic Agency Tawil is also the first Jewish boy to participate in the adoption tradition : Shlomo and Nehama Tawil , parents of seven boys, in 1993 wrote a letter the president asking for the honor Yair wrote a letter the president citing the 2009 decree and asking for the designation of godson The president her tweets and photos described 3.4 million Twitter followers a " magical moment " with a " marvelous family. She described Yair as " a total sweety ," and his mother a " Queen Esther .’

Table 4.11: An example showing similarity between agree and discuss labelled data in the FNC-1 corpus

relations and their components are an important bit of information for stance detection.

The error analysis shows that the computational results are highly affected by the unbalanced distribution of stance classes in the FNC-1 corpus. Besides having a balanced corpus, a solution to this problem could be using a three-stage architecture for stance detection. In the three-stage architecture, we can filter unrelated and discuss data at the first and second stages respectively. Finally, in the third stage, we can concentrate on minority classes agree and disagree implementing hand-curated class-specific features.

Chapter 5

Relating Source Expertise to Claim Credibility: an empirical study

5.1 Introduction

In this chapter, I discuss an empirical study where I conduct a survey to find the relation between readers' judgement on source expertise and claim credibility in news articles. The big picture for doing this study is to establish the dependency of claim credibility on the source expertise.

This study can be seen as a preliminary study for my next stage work where I enrich a pre-existing corpus with source expertise data. Previous works showed that speaker or user meta-data is useful in rumour stance classification (Aker et al. 2017, Dungs et al. 2018, Gorrell et al. 2019) and truth assessment of the information (Wang 2017, Long et al. 2017). My work is different from these because I work in long-text news documents. While judging the credibility of a claim, people might use the source expertise as an essential element. That might help them to decide the credibility of the news document and hence, help them judge the truthfulness of the news. Previous works (Dungs et al. 2018, Gorrell et al. 2019) have argued that stance information is useful to assess the truthfulness of a text. In the other direction, source expertise might also be helpful to establish the article body's stance towards the headline. We can establish this dependency as:

> source expertise –> stance –> truthfulness of text

Hovland & Weiss (1951) presented trustworthiness as an element of the credibility. I hypothesize that it might be the same in the case of texts or other communicated messages.. Thus, I believe that stance information might be useful to assess the credibility of the text. I hypothesize that if source expertise is useful to assess the claim credibility then it might be useful to detect the stance. With this hypothesis, I decided to assess the dependency of claim credibility judgement on the source expertise judgement. This preliminary study provide me with a foundation for an extensive data collection at the next step.

To my knowledge, this is the first work in collecting and using expertise level of sources that are explicitly mentioned in news documents for stance detection. Before deciding to collect source expertise data, I carry out a study to analyse whether there exists any dependency between a reader's judgement on source expertise and claim credibility. Thus, in this chapter, I deal with the following research question.

> RQ3: Is a reader's judgement of claim credibility positively correlated with his/her judgement of the level of expertise of the source who is making that claim?

Not just the presence of sources but also their domain expertise is equally important to decide whether the information they are conveying is credible to readers. A reader's understanding of the source's level of expertise could influence their opinion on the credibility of the article. Thus, expertise of a source could be an useful information to assess credibility of their claims that could be further useful in the stance detection. To discuss the problem in detail, let us consider an example given in Table 5.1.

In the example given in Table 5.1, there are two different sources, **the 21-year-old boy** and **a Perth arachnid expert Dr Volker Framenau** who are making conflicting claims. Now, on what basis a reader could decide which information to consider true? The source's level of expertise can be one of the parameters to decide on that. If an expert source is claiming something

Headline	Tropical spider burrowed under man's skin through appendix scar and lived there for THREE DAYS
News article	The 21-year-old was on his first trip to Bali. **He** told News Corp yesterday *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin.* There's just one problem. *Spiders*, according to **Perth arachnid expert Dr Volker Framenau**, *don't burrow in skin.* *"They don't have the tools, the armature, to do this sort of stuff,"* **Dr Framenau** said.

Table 5.1: A news document with **sources** and *claims*

related to his/her domain expertise then the reader might consider such information to be credible. Considering this hypothesis, the reader might infer that **Dr Framenau** is giving credible information because he is an arachnid expert.

In news reporting, several entities can be considered as sources such as a news link, a news writer, an eye-witness or some reports that provide information for news reporting. In this work, I followed the definition of the source given in Chapter 3 where a source is defined as a communicative agent or artefact to which contents are attributed. The source can be a person, an organization or even a report. For my study, I considered only those sources that are explicitly mentioned in news documents. In some cases, sources are implicit such as in passive voice sentences. My study doesn't include such implicit sources. For instance,

> Seth Rogen is said to be in discussions to play Jobs' colleague and Apple co-founder Steve Wozniak, but no official announcements have been made.

Here, I deal with the expertise of sources and the credibility of their respective claims. Reich (2011) defined source credibility as the degree to which the information from the source is perceived as accurate, fair, unbiased and trustworthy. A source believable to a reader may not be the same for another (Wathen & Burkell 2002, Metzger et al. 2010). The credibility of the source varies among people as per their belief systems (Metzger et al.

2010). Some people may find a source trustworthy but others may not agree. Thus, we can say that source credibility and trustworthiness are subjective concepts.

Hovland et al. (1953) presented trustworthiness and expertise as the fundamental components of the source credibility. Wertgen & Richter (2020) used source expertise as a factor to rate the credibility of the source. It means the expertise of sources adds value to their credibility. The source credibility is an abstract concept that varies as per what a person believes. However, the source expertise can be judged based on features like the source's qualification, title, working years, associated organization etc. Thus, I decided to collect people's ratings on the source expertise. Additionally, I asked survey participants to judge credibility of different claims made by sources such that I can analyse if there exists any correlation between the source expertise and claim credibility.

I did a quantitative study to collect readers' judgements on the source's level of expertise and claim credibility, and did statistical analysis of collected data to validate whether my presumption about source expertise and claim credibility is true. I define the null and alternative hypotheses for this study as follows:

> H_0: There is no significant relationship between the readers' judgements on source expertise and their judgements on claim credibility.
>
> H_a: Readers' judgements on claim credibility are positively correlated with their judgement of source's level of expertise.

5.2 Literature Review

One of the early studies on credibility by Hovland & Weiss (1951) presented trustworthiness as an element of the source credibility. Hovland & Weiss (1951) argued that the perceived trustworthiness of sources highly affects opinions on the communicated information. Hovland & Weiss's (1951) work also showed that the fairness and justifiability judgements of the communicated information rely on the source credibility. The sources with low credibility were found to be less fair and less justified than the source with high credibility. Later, Hovland et al. (1953) defined the source credibility as the combined effect of the expertise and trustworthiness of the source. Giffin (1967) argued that the trustworthiness of the speaker depends on the listener's perception of the five characteristics of the speaker that are:

- Expertness
- Reliability
- Intention
- Dynamism
- Personal attraction

The Collins English Dictionary defines credibility, trustworthiness and expertise as follows.[1]

- Credibility: The quality of being believed or trusted
- Trustworthiness: Worthy of being trusted; honest, reliable, or dependable
- Expertise: Special skill, knowledge, or judgment; expertness

A similar work by Canini et al. (2011) argued that the perceived credibility of a person affects other people's judgments on used car values. Their study showed that an expert Twitter user's opinions on car price can affect the judgement of other people on the used car values. It means people make decisions based on information that an expert communicates (Canini et al. 2011). If there are claims coming from highly credible people then they are considered as more plausible than the ones coming from less credible people (Wertgen & Richter 2020). In Wertgen & Richter (2020)'s work, people rated the source credibility depending entirely on the source's level of expertise. Wertgen & Richter (2020) considered the source expertise as the explicit

[1] `https://www.collinsdictionary.com/dictionary/english`

statements about how much expertise a person possesses in a field and other expertise-related information such as the profession, occupation or an academic title. For the study, Wertgen & Richter (2020) used 36 short stories related to everyday situations like vacations, restaurant visit etc where each story consists of eight sentences. The following is an example from Wertgen & Richter (2020).

- Low expertise:

  Sandra had almost no knowledge about astronomy and stars.

- High expertise:

  Sandra had a lot of knowledge about astronomy and stars.

My study is different from Wertgen & Richter (2020)'s work as my objective is to analyse the dependency of the perceived claim credibility on the perceived source expertise. Furthermore, I am working on news documents and considering sources that are components of attribution relations.

Roberts (2010) showed that there exist a high correlation between the scales used by Meyer (1988) and Flanagin & Metzger (2000) for the messenger and message credibility respectively. Thus, Roberts (2010) suggested that those scales could be used to measure the messenger and message credibility. Furthermore, Roberts (2010) argued that there exists a conceptual overlap of credibility among the source, message and medium of transmission. For this reason, I used the same scales to rate source expertise and claim credibility in my survey. The following sub-section describes literature related to rating scales that I used in my study.

Following arguments from Hovland et al. (1953), Canini et al. (2011) and Wertgen & Richter (2020), I decided to consider source's level of expertise as a parameter to decide their credibility. Therefore, I asked participants in the survey to rate the expertise level of sources that are explicitly mentioned in news articles.

5.2.1 Source expertise rating

There are many studies involving source credibility and expertise where different rating scales were used. Reich (2011) used a 6-point scale measure to rank the credibility of sources that are as follows:

1: Highly credible

2: Credible

3: Fairly credible

4: Not very credible

5: Not credible

6: Not credible at all

Another work by Wertgen & Richter (2020) used a 7-point scale to rate source credibility based on their expertise. The 7-point scale ranges from 1 representing *not credible at all* to 7 which represents *very credible*. Revilla et al. (2014) argued that a 5-point rating scale yields more high-quality data than a 7-point or higher rating scale in agree-disagree questions.

I opt to use a 5-point rating scale for source expertise based on work by Wertgen & Richter (2020) and Revilla et al. (2014). There is also an additional option for the indecisiveness that is based on pilot studies discussed in Section 5.3. I added a sixth option of indecisiveness in the rating scale as per outcomes of the pilot studies. The source expertise rating scales are as follows.

0: Not an expert

1: Barely an expert

2: Moderately expert

3: Fairly expert

4: Highly expert

Can't decide the source's level of expertise

5.2.2 Claim credibility rating

There are different types of scales used to rate the credibility of information or message conveyed by a source. Meyer (1988) introduced a 5-item credibility criteria for newspaper (any messenger or source) scoring based on 12 credibility factors. Those 5-item credibility criteria are as follows (Meyer 1988).

- fair/unfair
- biased/unbiased
- tells the whole story/does not tell the whole story
- accurate/inaccurate
- can be trusted/can't be trusted

Here, Meyer (1988) collected a true/false response for each above-mentioned credibility factors.

Flanagin & Metzger (2000) used a similar five credibility criteria to rate the message that are as follows.

- believable
- accurate
- trustworthy
- biased
- complete

For each of the given criteria, they used a 7-point Likert scale. The scale ranges from 1 representing "not at all" to 7 which represents "extremely".

Gupta & Kumaraguru (2012) used several options during annotation to assess the presence of credible information in each tweet. If the tweet is related to a specific event then the credibility of information is rated as any one of the following four options.

- Definitely Credible
- Seems Credible
- Definitely Incredible
- I can't Decide

Shin et al. (2019) measured the credibility of a document on a binary scale, 0 as not credible and 1 as credible.

As the pilot study discussed in Section 5.3 and Roberts (2010)'s argument that there exist a high correlation between the messenger and message credibility scales, I used a similar rating scale for the claim credibility as for the source expertise. Besides rating scales, previous works (Meyer 1988, Flanagin & Metzger 2000) have also used credibility criteria to rate the information. I didn't define any criteria for the claim credibility rating because my objective is to study the influence of source expertise judgement on the credibility (believability or trustworthiness) of their claims. Similar to source expertise rating scale, I added an option for the participant's state of indecisiveness. I used the following rating scale for the claim credibility judgement in my study.

0: Not credible at all

1: Barely credible

2: Moderately credible

3: Fairly credible

4: Highly credible

Can't decide the claim's credibility

5.3 Pilot

To ensure the usability of the survey, the appropriateness of rating scales and to validate whether participants find enough information to rate source expertise and claim credibility in the survey, I did two pilot studies. I did the first pilot at the preliminary stage and the second before the final survey. The purpose of those pilots is also to ensure allocated time and effectiveness of guidelines in the survey. In the first pilot study, nine participants judged expertise within two categories.

- Expert
- Non-expert

Additionally, participants also judged whether the source's claim is credible or not with following two categories.

- Yes
- No

I conducted this pilot study online within the University during an NLP group meeting. The participants include 4 academics and 5 post-graduate research students from the Open University. As it was my preliminary work, I used Microsoft Forms[2] to prepare my survey questions. I provided a detailed annotation guideline of two pages to everyone that is given in Appendix B. I shared the annotation guideline as a Microsoft word document during the study. There were 5 news bodies in separate forms for which questions for source expertise and claim credibility were asked. In total, there were 48 attribution relation in those 5 news bodies. Thus, forms contain 48 questions each for source expertise and claim credibility annotations. I shared links of each Microsoft form to the participants through Microsoft Teams. The total time allocated for reading guidelines and annotating all forms was 26 minutes.

The pilot participants mostly commented on the unfit nature of rating scales

[2] https://forms.office.com/

for both source expertise and claim credibility. Participants suggested that instead of binary rating scales, multi-level scales could be effective to make the right judgement.

Therefore, during the second pilot study, following Reich (2011)'s work I decided to opt for a 5-point scale for both source expertise and claim credibility judgements. The source expertise rating scale is as follows.

0: Not an expert

1: Barely an expert

2: Moderately expert

3: Fairly expert

4: Highly expert

During this second pilot, I shared a survey with two participants where I asked them to rate the source's level of expertise and claim credibility on a 5-point scale from 0 to 4. The participants in this testing were both post-graduate research students. I also requested them to provide comments if they find anything inconvenient during the study. I used the same format, guidelines and duration for this testing as for my final study, but with different news documents. Both participants in the testing commented that sometimes it is difficult to decide the source's level of expertise even though their related information is available in the article. Following are the examples that illustrate such indecisive situations. In the following examples, participants judged **source**'s level of expertise and *claim*'s credibility.

Example 5.1.
"Her presence was both overwhelming and comforting," **he (A supposed Catholic priest, Father John Micheal O'Neal)** said. *"She had a soft and soothing voice and her presence was as reassuring as a mother's embrace. The fact that God is a Holy Mother instead of a Holy Father doesn't disturb me."*

As per participants' response in Example 5.1, the expertise of the source is difficult to decide due to two reasons. First, the person reading the text may or may not believe in God. The person judging the text can be an atheist. Second, the presence of the word *supposed* has made the priest doubtful. It shows that a doubtful cue present in a source makes the source as well as their claims dubious.

> **Example 5.2.**
> **He (The 21-year-old Dylan Thomas)** told News Corp yesterday *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin.*

As per participants' response in Example 5.2, the source seems both expert and non-expert at the same time. The source could be judged as a non-expert because he is not an expert to confirm what creature the doctors pulled out from his skin. It could be a spider or any other creature. In contrast, some may judge the source as an expert because he is the one who experienced and eye-witnessed the incident.

Thus, to address such situations of indecisiveness I included an additional option for indecisiveness in the rating scales. While rating the source's level of expertise, even I don't know how expert is the Catholic priest (Example 5.1) in the context of presenting a fact about the God being Holy Mother. So, I opt to decide an additional option *Can't decide the source's level of expertise* besides other expertise scale from *0: Not an expert* to *4: Highly expert*.

Similarly, participants also mentioned about not being able to decide credibility of claims in some cases. As we saw in Example 5.2, the 21-year-old Dylan Thomas might be an eye-witness of the incident however, it is difficult to decide credibility of what he is claiming. So, I opted to add *Can't decide the claim's credibility* option in rating scales for claims.

5.4 Method

5.4.1 Estimating the sample size

In general, probability sampling of the population is a preferred method in the surveys because the random selection of the population reduces the sampling bias. Sampling bias occurs when some members of a population are systematically more likely to be selected in a sample than others. However, in my study, I adopted non-probability sampling because of constraints such as convenience and voluntary self-selection of the participants.

The size of the sample is estimated as 25 using a sample size calculator. [3] I gave the following inputs for the sample size estimation.

Test family = Correlation

Number of tails = One

Correlation co-efficient = 0.5

Significance level = 0.05

Power = 0.8

Here, *number of tails* is considered as *one* because as per my hypothesis the reader's judgement on claim credibility could be positively dependent on their judgment of source's level of expertise. The significance level is the probability of falsely rejecting the null hypothesis, and the most commonly used significance value is 0.05. Statistical power is the ability of the study to detect a result that exists in nature. Generally, a high statistical power is considered desirable. However, setting it too high may result in a sample size that is not practical. A value of 0.8 is often used in practice. Using all these values, the estimated population size is 25.

[3] `https://www.ai-therapy.com/psychology-statistics/sample-size-calculator`

5.4.2 Data selection for the survey

I decided to choose four news bodies from the training set of FNC-1 corpus [4] for the survey that is explained in Section 4.2. I limited the number of news bodies to 4 such that time spent on the survey will be no longer than 25 minutes. I didn't lessen the number of news bodies because I needed to collect data that is sufficient for the hypothesis testing. The length of each selected news articles is not more than 250 tokens. For the survey, I chose news bodies that have the following characteristics:

i. It should contain claims from both expert and non-expert sources.

ii. It should contain claims with some disagreements.

iii. It should not contain politically sensitive data that may cause discomfort for the participants to rate.

iv. All news bodies should be from the FNC-1 corpus.

Besides the above-mentioned features, I considered additional criteria for the new body selection which are the number of headlines associated with the articles and the presence of the stance label disagree. These features are considered such that the survey could yield a large number of data by associating each news body to more headlines for a computational work in future. Selecting a news body associated with a large number of headlines in the FNC-1 corpus can yield more data for the stance detection.

I took four news documents from the FNC-1 corpus for the survey. Appendix C contains four news documents that I used in my study. The FNC-1 corpus is derived from the Emergent corpus by Ferreira & Vlachos (2016). The headlines in the FNC-1 corpus are summaries of the news bodies which are mixed and matched with different news bodies. The Emergent corpus has 300 claims and around 2.5K headline-body pairs collected during a rumour debunking project (Silverman 2015).

In the survey, I didn't use headlines of news articles. I excluded the headlines from the survey content for the following reasons:

[4] `https://github.com/FakeNewsChallenge/fnc-1`

i. Generally, a headline is a representative text or the summary of a news article. The presence of a headline could make the readers biased to believe the claims related to the headline regardless of who made those claims, an expert or non-expert (Gabielkov et al. 2016, Dor 2003). That can directly affect my survey objective because the participant's entire judgements could rely on the headline text. That can highly impact the study if the headline contains information that aligns with the participant's belief. By just reading the headline, the participant could make judgements if the information in the headline confirms their pre-existing opinions. This phenomenon is also known as confirmation bias.

ii. The FNC-1 corpus has multiple headlines associated with the same news body. So, providing a single headline to a news body is not appropriate. It makes the data biased to a single headline. If I use the headline-body pairs in the survey, then such news bodies can be re-joined to other headlines in the corpus. It will make the data unsuitable for the second task of the stance detection. In the stance detection task, the news bodies are associated with their respective headline pairs as given in the FNC-1 corpus. Thereafter, I use the source expertise data collected in the survey for stance detection.

For the example in the survey guideline, I opt for a representative text that is short, not a part of the actual survey questionnaire, and that shows conflicting claims from two sources. Choosing a different news article helps to get rid of any possible bias that may occur due to early exposure to the actual survey contents. Following is the example in the guideline with **source** and *claim* to rate their respective expertise and credibility.

> **KLAS-TV in Las Vegas** is reporting *that Jose Canseco was accidentally shot in his home*. **A neighbor** tells 8 News NOW *former baseball star Jose Canseco was hurt in an accidental shooting Tuesday afternoon at his house on the eastside of the Las Vegas valley*. **Metro Police** confirm *there was an accidental shooting at the address*, but

> would not confirm *the former player was hurt*. However, **records** show *Canseco owns the home where the shooting happened*.

In the survey, I asked participants to judge and rate each claim that are explicitly made by different sources. The source's claims may not always be present in the same sentence. Claims may appear in different sentences, sometimes despite of them being parts of the same attribution relation. I represent such claims as different entities because of the following reasons.

i. If the source is making two claims, one specifically related to the news context but, another a general statement then, annotating both claims as a single claim could result in mixed credibility ratings. People may find the source highly expert while giving information directly related to their domain and context. However, it may not be true when the source makes general statements. For example:

   > [*"That's a professional skin-digger,"*]$_{claim1}$ **Perth arachnid expert Dr Volker Framenau** said. [*"There's a lot of nasty stuff out there."*]$_{claim2}$

   Here, the *claim1* is precisely related to the context where it discusses whether the creature pulled out from a boy's skin is a spider or not. However, *claim2* is a general statement that is presented as supporting evidence to *claim1*. Participants could have reservations to consider the same source as an expert for *claim2*. Thus, I choose to collect judgements for those two claims separately.

ii. There could be situations when the factual information and opinions should be treated differently during source expertise and claim credibility judgements. For example.

   > **Terrence Donilon, a spokesman for the Archbishop of Boston**, told Metro.co.uk [*they had no record of O'Neal being a priest*]$_{claim1}$. [*"We do not have a priest of this name. I believe this could be a hoax story."*]$_{claim2}$

   Here, the *claim1* is a fact presented by the source. However, *claim2* is a mixture of factual information and the source's own opinion. The *claim2* can also be seen as supporting evidence provided for *claim1*. Therefore, *claim1* and *claim2* could be treated differently for the credibility judgements.

5.4.3 Question randomization

Goodhue & Loiacono (2002) argued that although participants may find intermixed questions more confusing than the grouped questions, combining the questions makes the measures more reliable and error terms less correlated than the grouped questions. While ordering questions in a survey, general questions should appear before the questions with specific contents (McFarland 1981). However, highly-specific questions are not affected by the question order.

I have two types of questions asked for the same news text, one related to the source's expertise and another related to the claim credibility. There are two broad groups of questions related to the source expertise and the claim credibility. Thus, I prepared two subsets referring to each group of questions and implemented the in-built question randomisation available Qualtrics[5]. The Qualtrics is a web-based software that I used to create my survey and to generate reports. Although all questions are randomised, there is always a source expertise question followed by a claim credibility question.

Following the professional practice, my empirical study is reviewed by the Open University Human Research Ethics Committee (HREC). My study received a favourable opinion from the HREC with a reference number of 3924.

[5] `https://www.qualtrics.com/uk/lp/surveys/`

5.5 Results

In this section, I summarise the data collected in the survey and a statistical test that I did to check whether the null hypothesis is accepted or not.

5.5.1 Data in the Survey

Table 5.2 shows the number of news articles and questions included in the survey for data collection.

	News 1	News 2	News 3	News 4
Source expertise questions	7	6	6	5
Claim credibility questions	7	6	6	5

Table 5.2: Survey question distribution

From the Table 5.2 we can say,

Total source expertise questions = 24

Total claim credibility questions = 24

25 people participated in the survey. Let us represent the reader's judgement on the source expertise and claim credibility by the variables X and Y respectively. X and Y can be represented as follows.

X = {x_{ij} where i=[1,25] and j=[1,24]}

Y = {y_{ij} where i=[1,25] and j=[1,24]}

Here, i represents the survey participants and j represents the survey questions.

The data collected in the survey is visually represented in Figure 5.1 and Figure 5.2 representing respective source expertise and claim credibility data distribution of the study. There are 1200 data points collected in the

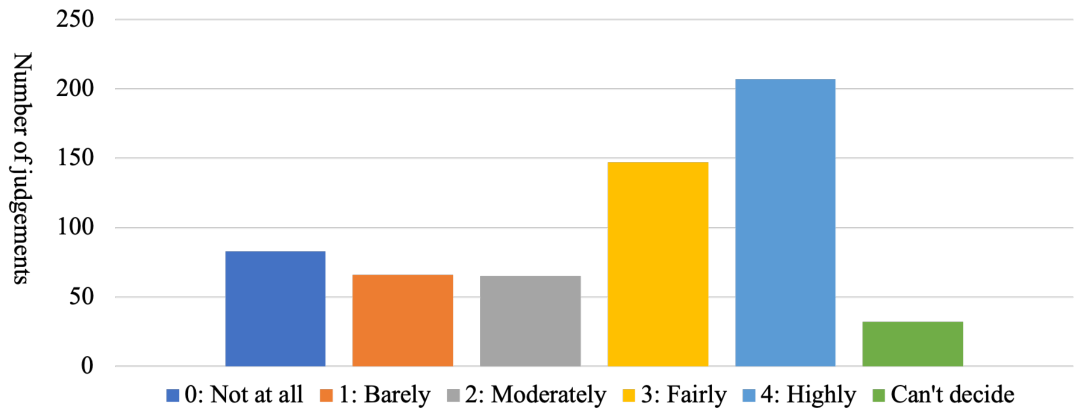


Figure 5.1: Source Expertise data distribution

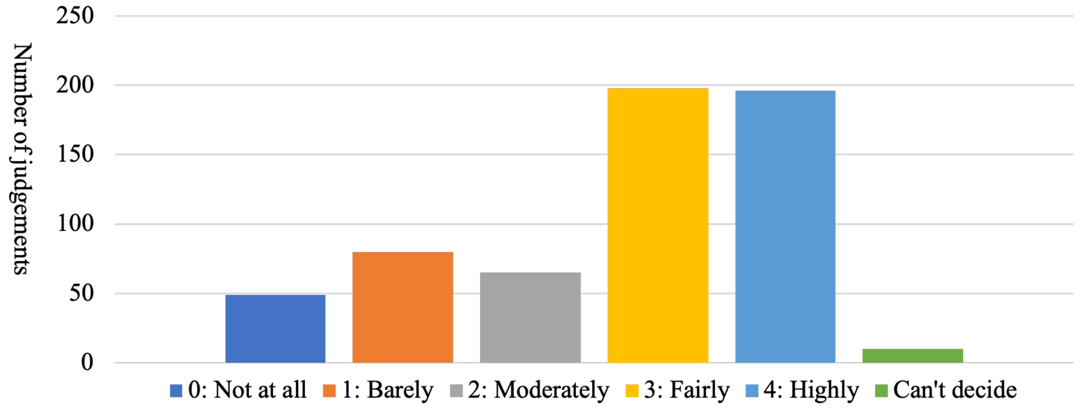


Figure 5.2: Claim credibility data distribution

survey.[6]

5.5.2 *Can't Decide* category data analysis

There are cases where annotators selected the *can't decide* category. For convenience, I call that category the *CD*. I observed that 38 questions where at least one of the source expertise or claim credibility judgement is rated as a *CD*.

Table 5.3 shows how often readers were unable to decide source expertise and claim credibility in news articles. The table shows that readers are more indecisive to judge source expertise than claim credibility. This result is different from what I previously presumed. I had hypothesised that participants would be more confident to rate source expertise than to rate

[6] https://figshare.com/s/9a90210bb9450ecf7395

		Source Expertise Judgements	
		CD	**0-4**
Claim Credibility Judgements	***CD***	6	6
	0-4	26	

Table 5.3: *CD* occurrence in data

claim credibility because of source-related information present in news articles. However, the data shows different results. It could be because participants missed considering all source-related information present in news articles. Another reason could be that sources in the survey may contain such information that opposes participants' existing belief systems. For instance, if participants are theist and if they read questions with text like "A supposed Catholic priest (Father John Micheal O'Neal)'s claims of seeing God as a woman when he died for 48 minutes are being described as a hoax.", it could leave them puzzled and indecisive.

There are two survey questions that have the highest *CD* rating (6 out of 25 ratings) for source expertise judgement that are as follows.

- **A supposed Catholic priest (Father John Micheal O'Neal)**'s claims *of seeing God as a woman when he died for 48 minutes* are being described as a hoax.
- *"Her presence was both overwhelming and comforting,"* **he (A supposed Catholic priest, Father John Micheal O'Neal)** said. "She had a soft and soothing voice and her presence was as reassuring as a mother's embrace. The fact that God is a Holy Mother instead of a Holy Father doesn't disturb me."

The given texts extracted from the survey shows that survey participants are indecisive to rate the same source's level of expertise. It could be because the priest is introduced with a hedging word *supposed* that makes the person doubtful.

The claim credibility question that has highest *CD* rating (3 out of 25 ratings) is as follows.

> "Her presence was both overwhelming and comforting," **he (A supposed Catholic priest, Father John Micheal O'Neal)** said. *"She had a soft and soothing voice and her presence was as reassuring as a mother's embrace. The fact that God is a Holy Mother instead of a Holy Father doesn't disturb me."*

Here, we can see that the same question text as source expertise has the highest number of indecisive situations for claim credibility judgement. It could be because indecisiveness about source expertise affected the judgement decision for claim credibility.

5.5.3 Hypothesis testing: Chi-square test of independence

In this section, I used Chi-square test for hypothesis testing. Chi-square test of independence tests the statistical independence or association between two or more categorical variables (Zibran 2007, McHugh 2013). As defined Section 5.1, null and alternative hypotheses are,

> H_0: There is no significant relationship between the readers' judgements on source expertise and their judgements on claim credibility.
>
> H_a: Readers' judgements on claim credibility are positively correlated on their judgement of source's level of expertise.

Although the rating scale from 0 to 4 in my study is ordinal, including the category CD makes my data categorical. So, I implemented a non-parametric test for hypothesis testing.

Table 5.4 represents the contingency table that summarises the relationship between source expertise judgement and claim credibility judgement. In Table 5.4, we can see that the diagonal values are higher than the rest. This depicts that the claim credibility increases with the source expertise. The degree of freedom(dof) refers to the maximum number of logically independent values that have the freedom to vary in the data sample. There are 6 rows and 6 columns in the Table 5.4. Thus, dof is calculated as product

		Source expertise judgement						
		0	**1**	**2**	**3**	**4**	**CD**	**Sum**
Claim Credibility Judgement	**0**	**37**	6	0	2	0	4	49
	1	31	**36**	6	2	1	4	80
	2	8	11	**23**	11	6	6	65
	3	3	9	28	**100**	51	7	198
	4	4	4	6	28	**149**	5	196
	CD	0	0	2	4	0	6	10
	Sum	83	66	65	147	207	32	

Table 5.4: Contingency table for Chi-square test

of one less row and column counts that is as follows.

$$dof = 25$$

The test statistics and p-value of the Table 5.4 is computed as follows.[7]

$$\chi^2 = 680.401$$

$$p = 1.763e - 127$$

A critical value is a value that can be compared with the test statistic to indicate if the null hypothesis can be rejected. In the given context, let us consider

$$\alpha = 0.05$$

It means a probability of 95% is used, suggesting that the finding of the test is quite likely given the null hypothesis of the test that the variables are independent. If the test statistic is less than or equal to the critical value, we fail to reject the null hypothesis.

The critical value for the given data distribution is computed as follows.[8]

$$critical\ value = 37.652$$

In the given context,

$$\chi^2 > critical\ value$$

Additionally, comparing p-value with the significance level, it is found that

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html

$$p < \alpha$$

A Chi-Square Test of Independence is performed to assess the relationship between a reader's judgement on the source's level of expertise and claim credibility. There is a significant relationship between the two variables,

$$\chi^2(25, N = 600) = 680.40, p = 1.763e - 127$$

A reader's judgement on claim credibility is dependent on the judgement of source's level of expertise who is making that claim.

5.5.4 Hypothesis testing: Pearson's correlation test

Pearson's correlation test is a parametric test suitable to test hypotheses in continuous data. Norman (2010) argued that parametric tests like Pearson's correlation can be a robust test choice for ordinal data like ones collected using Likert scale. In my study, both source expertise and claim credibility have ordinal rating scales from *0:Not an expert* to *4:Highly expert*. However, the sixth rating scale for both judgements is a different category *CD*. So, I decided to replace the category *CD* by *6* for the correlation calculation. Finally, I test whether claim credibility judgement is correlated to the source expertise judgement.

Correlation test checks whether two variables are related without assuming cause-and-effect relationships. Pearson's correlation is the calculation of the covariance (or expected difference of observations from the mean) between the two variables normalized by the variance or spread of both variables. The covariance (covr) is calculated as follows:

$$covr = \frac{\sum[(x - mean(X)(y - mean(Y))]}{n - 1} \tag{5.1}$$

The sign of the covariance can be used to interpret whether the two variables X and Y change in the same direction (represented by a positive value) or in different directions (represented by a negative value). A covariance value of zero indicates that the variables are independent.

Pearson's correlation coefficient(PCC) is calculated as the following.[9]

$$PCC = \frac{covr(X,Y)}{sd(X) * sd(Y)} \tag{5.2}$$

Here, sd(X) and sd(Y) represent standard deviations of the variables X and Y respectively.

Considering X as source expertise judgement and Y as claim credibility judgement, we obtained the following results.

$$PCC = 0.651$$

$$p = 6.857e - 74$$

$$\alpha = 0.05$$

Here,

$$p < \alpha$$

Thus, the null hypothesis is rejected. It suggests that there exists a linear relationship between readers' judgement on source expertise and claim credibility.

5.6 Conclusion

The purpose of this study is to gain a better understanding of the association between the source's level of expertise and their claim credibility in news documents. The statistical test suggests that the presence of highly expert sources means the presence of more credible claims in a news document. The results of my study show that there exists a strong association between source expertise and claim credibility judgements.

As the statistical analysis of this empirical study suggests that claim credibility judgements are dependent on source expertise judgements, I collect more source expertise data in the next stage for computational purposes.

[9] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html`

Collecting source expertise data on an existing stance annotated corpus allows me to study the role of the source's level of expertise in the stance detection task.

Chapter 6

Role of source expertise in stance detection

6.1 Introduction

The previous Chapter 5 concluded that the reader might use the expertise of the source to judge the credibility of the claim. On the basis of that outcome, I collect more source expertise data and study its role to detect stance of news articles in this chapter. So far,to my knowledge, no one has explored the role of source expertise in deciding the stance of news articles. The research question that I intend to answer in this chapter is as follows.

> RQ4: What is the role of source expertise information in detecting the stance of a news body to its headline?

To the best of my knowledge, there is no corpus with news articles available that has both source expertise and stance annotations. So, I decided to add source expertise information as an extension to an existing corpus that has stance annotation. For this work, I used a subset of FNC-1 corpus that is explained in Section 4.2. So, I am using a subset of the same corpus to collect source expertise data. Once I have source expertise and stance annotated data, I can implement a machine learning model to assess how useful source expertise information is to detect stance.

For source expertise data collection, I followed the same selection criteria for news articles and rating scales that are respectively detailed in Section 5.4.2 and Section 5.2.1 of Chapter 5. The source expertise rating scales used in the data collection are as follows. Rating scales from 0 to 4 are at an increasing order of expertise of sources.

0: Not an expert

1: Barely an expert

2: Moderately expert

3: Fairly expert

4: Highly expert

Can't decide the source's level of expertise

I used a crowd sourcing platform for data collection. I opted to use such platform because of number of people required to annotate data in 20 different surveys. Questions in each survey were annotated by three different annotators. In this chapter I describe three activities. Firstly, I ran a pilot with two different crowd sourcing platforms: Amazon Mechanical Turk (AMT)[1] and Prolific[2]. This allowed me to assess platform reliability and also to analyse data reliability in crowd sourcing platforms. Section 6.3 contains details of the pilot. Secondly, I collected source expertise data using Qualtrics survey.[3] Details of the data collection is in Section 6.5. Thirdly, I used those source expertise data for the stance detection task that is described in Section 6.7.

Through this work, I contribute to enrich a corpus with source expertise data that I call the *FNC-SE* corpus. The corpus can be further utilised for the stance detection. Additionally, I also assess the role of source expertise to computationally detect the stance of news articles to their headlines. So far, to my knowledge, my work is the first to explore the usefulness of the source's level of expertise to detect stance.

[1] `https://www.mturk.com/`

[2] `https://www.prolific.co/`

[3] `https://www.qualtrics.com/uk/lp/surveys/`

6.2 Literature Review

Previous studies (Wang 2017, Long et al. 2017) showed that speaker meta-data is useful to assess the truthfulness of information. Many works (Derczynski et al. 2017, Gorrell et al. 2019, Dungs et al. 2018) considered stance information an essential feature to predict the veracity of a rumour. Zubiaga et al. (2018) proposed a rumour classification architecture where the stance classification is presented as a preceding step for the veracity classification of the rumour. Considering those, I anticipated that the **source** meta-data might be useful in detecting the stance. An early work of using speaker features in fake news detection by Wang (2017) used speaker meta-data like party affiliation, current job, home state and credit history. In the credit history Wang (2017) included the historical counts of inaccurate statements for each speaker. For instance,

> **Example 6.1.**
> Truthfulness Label: *Barely true*
> Statement: *Says the paperback edition of Mitt Romneys book deleted line that Massachusetts individual mandate should be the model for the country*
> Speaker: *Rick Perry*
> Speaker's job title: *Governor*
> State info: *Texas*
> Party affiliation: *Republican*
> Credit History: [barely true counts, false counts, half true counts, mostly true counts, pants on fire counts] = *[30, 30, 42, 23, 18]*

Wang (2017) created the LIAR dataset containing 12.8 thousand short statements collected from politifact labelled with six different labels of truthfullness- pants-fire, false, barelytrue, half-true, mostly-true, and true. Additionally, Wang (2017) showed that speaker related meta-data enhance the performance of fake news detection. Another work by Long et al. (2017) used the LIAR dataset (Wang 2017) and showed that speaker profiles provide valuable information for fake news detection. Long et al. (2017)

implemented an attention-based long short-term memory network model to get the state-of-the-art results in the LIAR dataset.

For the stance detection in news articles, the role of speaker meta-data is not explored so far to the best of my knowledge. However, works like Aker et al. (2017) explored twitter user meta-data to study their usefulness in the rumour stance classification . Aker et al. (2017) used user features like whether the user is a verified Twitter user, their number of followers and following information, whether users provided any description about themselves and made their geo- locations available. The user meta-data is also explored in fake news spreader detection in the Twitter data(Rangel et al. 2020, Rath et al. 2022). An author profiling shared task by Rangel et al. (2020) involved using user's fake news sharing behaviour to determine if that user is fake news spreader or not.

Considering given literature review of a user's meta-data in fake news detection and stance detection, I decided to study and analyse the usefulness of source expertise to detect stance in long-texts like news articles.

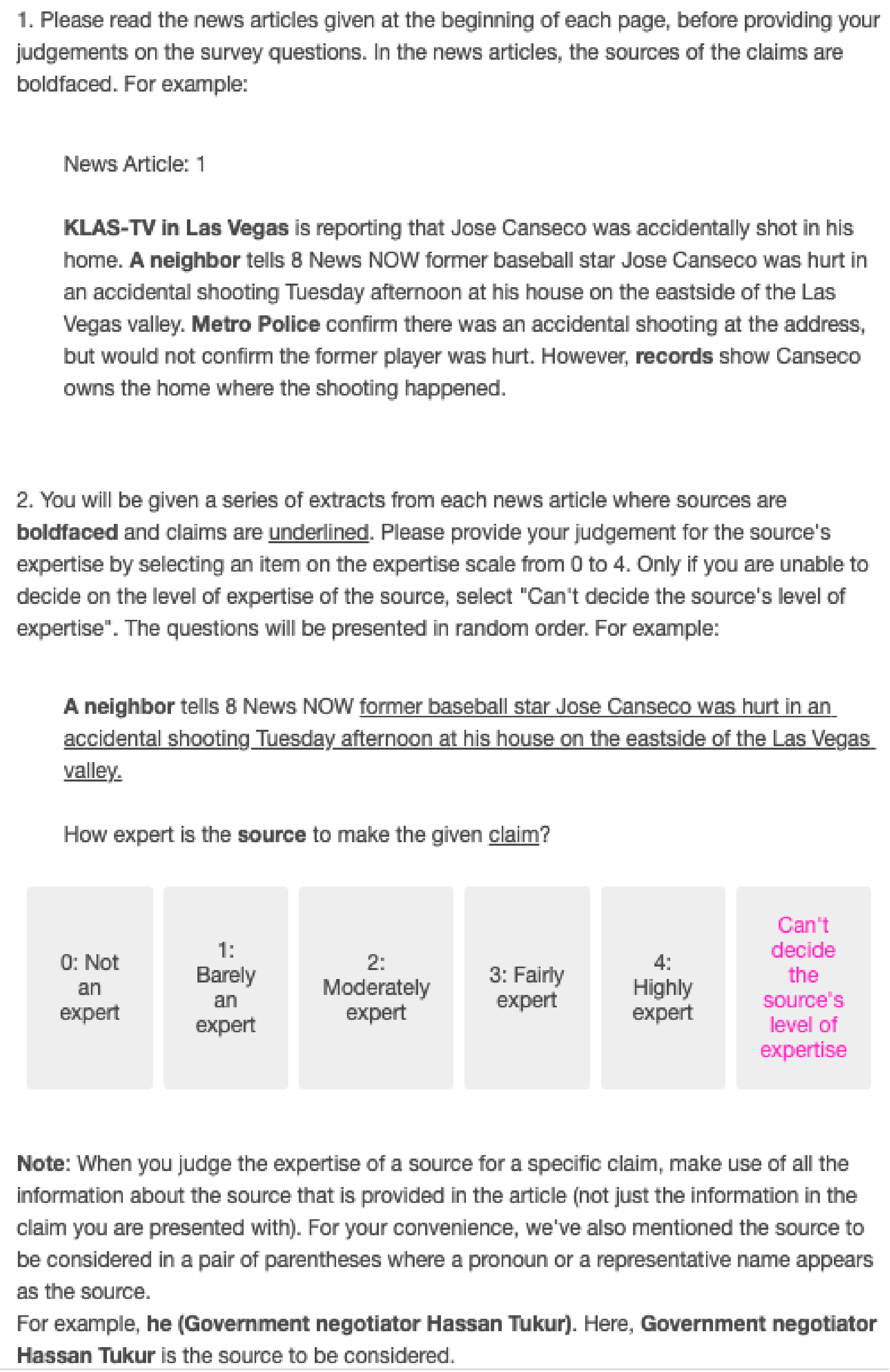
6.3 A pilot study

To check the quality of collected data, I decided to run a pilot on the two crowd-sourcing platforms Amazon Mechanical Turk[4] and Prolific[5]. Both of those platforms are widely used in the research with the Prolific being popular with the psychology related research in recent times. Thus, I wanted to assess which platform performs better for my work with a reliable data collection. I opt to test their reliability based on the inter-rater reliability of data collected on the respective platforms. The outcomes of this pilot study are as follows:

- Amazon mechanical Turk is a better platform for my data collection based on the agreement coefficient of data.

[4] https://www.mturk.com/

[5] https://www.prolific.co/



1. Please read the news articles given at the beginning of each page, before providing your judgements on the survey questions. In the news articles, the sources of the claims are boldfaced. For example:

News Article: 1

KLAS-TV in Las Vegas is reporting that Jose Canseco was accidentally shot in his home. **A neighbor** tells 8 News NOW former baseball star Jose Canseco was hurt in an accidental shooting Tuesday afternoon at his house on the eastside of the Las Vegas valley. **Metro Police** confirm there was an accidental shooting at the address, but would not confirm the former player was hurt. However, **records** show Canseco owns the home where the shooting happened.

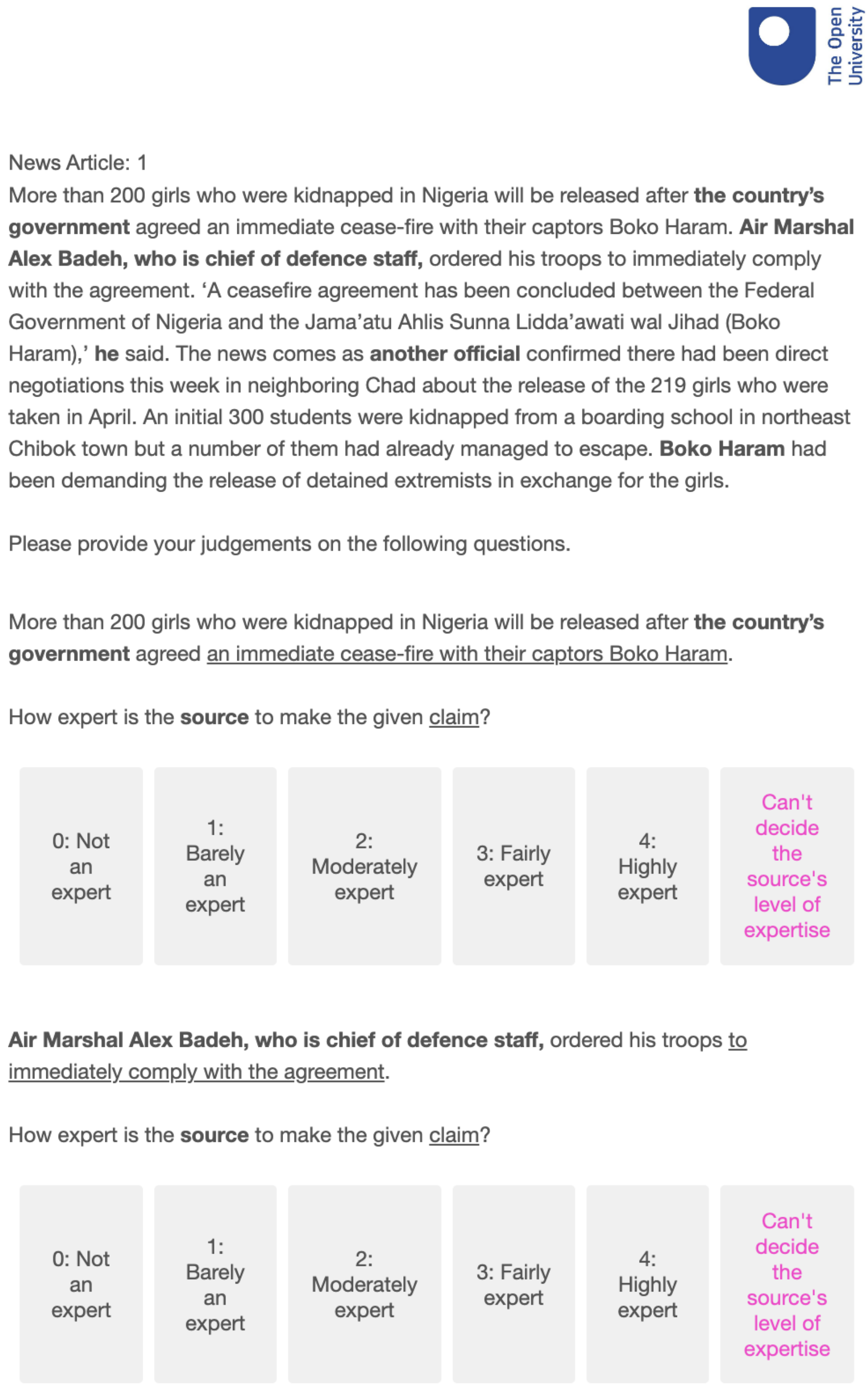
2. You will be given a series of extracts from each news article where sources are **boldfaced** and claims are underlined. Please provide your judgement for the source's expertise by selecting an item on the expertise scale from 0 to 4. Only if you are unable to decide on the level of expertise of the source, select "Can't decide the source's level of expertise". The questions will be presented in random order. For example:

A neighbor tells 8 News NOW former baseball star Jose Canseco was hurt in an accidental shooting Tuesday afternoon at his house on the eastside of the Las Vegas valley.

How expert is the **source** to make the given claim?

Note: When you judge the expertise of a source for a specific claim, make use of all the information about the source that is provided in the article (not just the information in the claim you are presented with). For your convenience, we've also mentioned the source to be considered in a pair of parentheses where a pronoun or a representative name appears as the source.
For example, **he (Government negotiator Hassan Tukur)**. Here, **Government negotiator Hassan Tukur** is the source to be considered.

Figure 6.1: Guidelines for source expertise annotation



News Article: 1

More than 200 girls who were kidnapped in Nigeria will be released after **the country's government** agreed an immediate cease-fire with their captors Boko Haram. **Air Marshal Alex Badeh, who is chief of defence staff,** ordered his troops to immediately comply with the agreement. 'A ceasefire agreement has been concluded between the Federal Government of Nigeria and the Jama'atu Ahlis Sunna Lidda'awati wal Jihad (Boko Haram),' **he** said. The news comes as **another official** confirmed there had been direct negotiations this week in neighboring Chad about the release of the 219 girls who were taken in April. An initial 300 students were kidnapped from a boarding school in northeast Chibok town but a number of them had already managed to escape. **Boko Haram** had been demanding the release of detained extremists in exchange for the girls.

Please provide your judgements on the following questions.

More than 200 girls who were kidnapped in Nigeria will be released after **the country's government** agreed an immediate cease-fire with their captors Boko Haram.

How expert is the **source** to make the given claim?

Air Marshal Alex Badeh, who is chief of defence staff, ordered his troops to immediately comply with the agreement.

How expert is the **source** to make the given claim?

Figure 6.2: A survey example in Mechanical Turk

- Weighted Gwet AC2 is the agreement coefficient that I should choose to assess the reliability of our data.

I used the same two surveys on both platforms to collect data. I used the software Qualtrics[6] to prepare my surveys for the pilot study. Figure 6.1 shows a screenshot of the annotation guidelines included in surveys for annotators. Figure 6.2 shows a screenshot from a survey conducted in the Mechanical Turk. We can see in the Figure 6.2 that **source**s in the news article are boldfaced. Additionally, claims from those **source**s are underlined in the questions. I did it on purpose so that annotators can find the required information conveniently. I had long news texts with questions so, I used this convention to save time such that each survey does not take more than 25 minutes. News articles are published in a plain text format without any **source**s boldfaced. However, my research makes it possible to highlight such information in plain-text news articles by detecting **source**, cue and *content* of an attribution relation as described in Chapter 3.

Table 6.1 shows annotation results of a survey that is annotated in the Amazon Mechanical Turk (AMT). There are twenty questions each of which is annotated by three annotators represented by Ant1, Ant2 and Ant3. In Table 6.1, we see that the category *4: Highly Expert* is often chosen by annotators. While doing annotation myself, I chose the category *4: Highly Expert* for 13 questions out of 20. In this example, we can see that data annotation is highly imbalanced and biased towards the category *4: Highly Expert*.

I used Gwet AC2 (Gwet 2014) statistic to compute the agreement among the raters. However, Kappa statistics are also a widely-used and accepted agreement coefficient that is used to assess the reliability of data. Kappa gives an insight into the annotation of the rare categories. This may cause loss of genuine agreement information and could not capture annotation of non-rare categories that may not happen by chance (Artstein & Poesio 2008). It becomes problematic especially when a particular category is selected

[6] `https://www.qualtrics.com/`

Question	Ant1	Ant2	Ant3
1	3	4	4
2	4	4	4
3	4	4	4
4	CD	CD	3
5	CD	4	4
6	4	4	4
7	4	4	4
8	4	4	3
9	4	4	4
10	3	4	4
11	CD	4	3
12	2	4	2
13	3	4	1
14	2	4	1
15	1	4	0
16	3	4	4
17	4	4	4
18	4	4	4
19	3	4	4
20	3	4	4

Table 6.1: A pilot survey data in AMT

more often than the remaining ones. The Kappa statistic is not appropriate to be used as agreement coefficient in this case because there are many sources that should be rated as the category *4: Highly Expert* considering my annotation. Here, the chance agreement may consider all or most of those annotations being made randomly by annotators, resulting in a high value of the chance agreement. This results in a very low kappa value of 0.044.

As shown in Table 6.1, annotators chose the category *4* very often making the data imbalanced. The kappa coefficient is not an ideal choice because it measures agreement on rare categories such as *0*, *1*, *CD* and misses information about the genuine agreement in non-rare categories such as *4* that might not happen by chance. To address such shortcomings, I used the AC2 agreement coefficient that gives an insight into the annotation of non-rare categories (Gwet 2014). AC2 has an assumption that only an unknown portion of the observed ratings are subject to randomness.

In my data, agreement and disagreement are not two different concepts because there exists a connection or structure among categories *0* to *4*. There is no disagreement between *not an expert* and *barely an expert*. Similarly, in the case of *fairly expert* and *highly expert* they are not completely different labels. For example, if an annotator annotates an item as *fairly expert* and another annotator label the same item as *highly expert* then those annotators are not disagreeing with each other. There is a partial agreement between those annotations because *fairly expert* and *highly expert* are not completely different concepts. Thus, I considered such partial agreement a part of the agreement coefficient calculation. Gwet (2014) showed that partial agreements among annotators can be captured using different weights while computing the agreement coefficient. Thus, I opted to implement different weightings to compute the agreement coefficient. I used the irrCAC library[7] in R to compute the agreement coefficient AC2.

	0	**1**	**2**	**3**	**4**	**CD**
0	1	0.9	0.7	0.4	0	0
1	0.9	1	0.9	0.7	0.4	0
2	0.7	0.9	1	0.9	0.7	0
3	0.4	0.7	0.9	1	0.9	0
4	0	0.4	0.7	0.9	1	0
CD	0	0	0	0	0	1

Table 6.2: Weight matrix used to compute agreement coefficient

Table 6.2 shows the weight matrix used to compute the agreement coefficient Gwet AC2. I considered ordinal weight distribution for category *0: Not an expert* to *4: Highly expert* because those categories are in an order. However, the difference between those categories is not known in the study. *CD* is the category which does not fall under the spectrum of source being an expert or non-expert. Rather, *CD* is more the state of annotators when they become indecisive about choosing a category for the source's level of expertise. The category *CD* has no connection with remaining five categories. Thus, only when two annotators choose the same *CD* category for an item, they are in an agreement or else there is a disagreement. The category *CD* is not considered for any partial agreements. Here, while using the weight matrix

[7] `https://rdrr.io/cran/irrCAC/`

given in Table 6.2, I obtained a Gwet AC2 coefficient of 0.71 that shows a satisfactory agreement among annotators.

A different way to assess the data reliability is to analyse how consistently annotators rate items. Amidei (2021) argued that annotators' rating consistency can be used together with the concept of an agreement to obtain a better assessment of the data reliability. In my data, it is difficult to do so. Firstly, it is not possible to get the same judge after a certain period to annotate the same data, as I used a crowd-sourcing platform (AMT) for data annotation. Secondly, to compare relative consistency among judges, I have not defined any criteria for annotator selection and any characteristics for their annotation comparisons.

While computing the Gwet AC2 coefficient for both surveys conducted in AMT and Prolific, I observed that agreement coefficients for both surveys are greater in AMT data (0.71 and 0.63) than the ones collected in Prolific (0.62 and 0.32). It suggests that AMT is more reliable than Prolific for my data annotations.

6.4 Assessing Reliability of attribution annotations

The FNC-1 corpus has only stance annotations. However, for my work I also need source expertise data along with attribution relations. So, first I extracted a subset from the FNC-1 corpus that I called FNC-SE. Then, I annotated attribution relations in the FNC-SE corpus following the annotation guidelines of the Political news attribution relation corpus.[8]

To check the reliability of the attribution annotation, around 10% of news articles used in the data collection are double annotated. For the second level annotation, I randomly selected 9 news documents among the ones used in surveys. A second annotator (one of my supervisors) annotated 9 news

[8] `https://github.com/networkdynamics/PolNeAR/blob/master/annotation-guidelines/guidelines.pdf`

articles from the FNC-SE corpus using the same annotation guidelines that I did. After the second level annotation of those nine documents, I manually analysed whether all three components of attribution relations **source**, cue and *content* spans match with the ones used in the survey. I observed that I annotated 51 attributions in those 9 news documents whereas the second annotator annotated 64 attributions in the same documents. However, in the case of the second annotation, there are 3 ARs the annotator was not certain about and 3 ARs that have no sources (passive voice texts). During the second-level annotation, I explicitly mentioned that despite whatever is mentioned in the guidelines, not to annotate ARs that have no explicit sources. I did it because the objective of the data collection is to collect the level of expertise for such sources that are explicitly mentioned in the news document. Therefore, I don't consider such AR annotations to calculate the agreement between annotators. It makes the total attributions annotated by the second annotator 58. I observed that 47 attributions are in common in annotations where all three **source**, cue and *content* spans matched.

To analyse the reliability of attribution annotation, I computed the agreement between two annotators. Here, annotators don't choose between labels. Rather, they identify if a relationship exists and if so, then they identify text spans that are part of the relation. For such situations, Wiebe et al. (2005) argued that the *agr* metric should be used, rather than the kappa statistic to find inter-annotator agreement. Here, we find the intersection of text spans identified as a part of the relation by both annotators. To find the *agr* metric, let us suppose two annotators a and b performed A and B annotations respectively. The *agr* is a directional measure of agreement that measures what proportion of annotation A that was also marked by annotator b. The *agr* metric is calculated as follows:

$$agr(a||b) = \frac{|A \cap B|}{|A|}$$

$$agr(b||a) = \frac{|A \cap B|}{|B|}$$

$$agr_{\text{ab}} = \frac{agr(a||b) + agr(b||a)}{2}$$

Here, $agr(a||b)$ measures agreement between a and b considering how many b's annotations match with A. Similarly, $agr(b||a)$ measures agreement between a and b considering how many a's annotations match with B. $|A \cap B|$ measures annotations (A) done by a that are also done by b. This means A matching with B. agr_{ab} measures the average of $agr(a||b)$ and $agr(b||a)$.

In the attribution annotation task, $agr(a||b)$ is 0.921 and $agr(b||a)$ is 0.810 resulting in an agreement coefficient agr_{ab} of 0.865. This agreement coefficient validates the reliability of attribution annotation in the data.

6.5 Data Collection in AMT

During this phase, I conducted 20 different surveys on Amazon Mechanical Turk (AMT). Three annotators annotated each survey. Each survey has four news articles extracted from the FNC-1 corpus. The total number of items (or questions) in each survey is between 24 and 30. We considered the following points while selecting news articles for those surveys:

- Number of tokens in each news article is 500 or less.
- Each news article must have at least one disagree stance while associated with headlines in the FNC-1 corpus.
- Four news articles in each survey cover different topics.

I restricted the number of news articles and their lengths such that each survey duration does not exceed 20 minutes. The reason for including news articles with disagree stance class is to get a balanced corpus at the end, unlike the FNC-1 corpus. Around 73% of the FNC-1 corpus is in the unrelated class.

Table 6.3 shows different information, like duration spent by three annotators for each survey, raw/observed agreement among annotators and the Gwet agreement coefficient values for 20 different surveys. The calculated average raw agreement and average Gwet AC2 agreement coefficient of 20

Data	Duration (secs)	Raw Agreement	Gwet AC2
1	692, 573, 528	0.422	0.719
2	455, 752, 697	0.205	0.225
3	253, 577, 1135	0.361	-0.070
4	475, 1153, 626	0.333	0.254
5	403, 1350, 537	0.367	0.517
6	572, 489, 382	0.344	0.730
7	474, 455, 716	0.322	0.653
8	378, 675, 280	0.298	0.182
9	606, 1245, 981	0.533	0.693
10	631, 1227, 1027	0.310	0.535
11	416, 626, 569	0.2	0.491
12	475, 475, 530	0.344	0.585
13	1050, 1128, 555	0.277	0.697
14	1097, 722, 577	0.540	0.757
15	347, 715, 333	0.377	0.674
16	968, 337, 1032	0.379	0.702
17	1012, 1127, 753	0.440	0.558
18	483, 650, 1171	0.172	0.493
19	685, 1005, 724	0.356	0.386
20	447, 1155, 555	0.3	0.608

Table 6.3: Agreement Coefficients in 20 surveys

surveys are as follows:

Average raw/observed agreement = 0.344

Average Gwet AC2 = 0.519

Following Landis & Koch (1977)'s benchmark scale for agreement values, 0.519 value of Gwet AC2 infers that we have moderate agreement among annotators. Despite subjective questions used in the survey that could be highly affected by the annotator's existing bias, the agreement coefficient assessment shows the data is reliable.

6.6 Enhancing FNC-SE corpus

In this section, I discuss enriching a subset of the FNC-1 corpus with source expertise annotations that can be further used in the stance detection. I selected 92 exclusive news documents from the FNC-1 corpus. The FNC-1 corpus is explained in Section 4.2. I enriched a subset of the FNC-1 corpus with source expertise information and called it FNC-SE corpus whose data collection details are given in Section 6.5. For the FNC-SE corpus, I collected source expertise data for each **source** in an AR of a given news body corresponding to the respective "Body ID" from the FNC-1 corpus. [9]

The FNC-SE corpus is created with 92 unique news bodies that are associated to their respective headlines in the FNC-1 corpus to form a complete news article. Those 92 unique news bodies consist a total of 645 attribution relations with each containing a source and its expertise level. While combining those 92 unique news bodies with their respective headlines in the FNC-1 corpus, we finally have 3,574 news articles in the FNC-SE corpus. I divided the FNC-SE corpus into training and test sets as per the article's Body IDs in the original FNC-1 corpus. I divided the FNC-SE corpus that way because FNC-1 corpus doesn't have any common news articles between training and test sets. It is good to have the training and test sets with completely different data such that the system is robust and can be used for different data. Therefore, articles in training and test sets of the FNC-SE corpus are different. Amongst 92 new articles, the training set has 72 and the test set has 20 news articles.

In the FNC-1 corpus, the same article appears in different stance classes as per the headline it is associated with. Therefore, I extracted headlines and stance labels corresponding to news articles used in our surveys. The distribution of the four stance classes in training and set of FNC-SE corpus is shown in Table 6.4.

	Agree	Disagree	Discuss	Unrelated
Training set	441	290	469	1513
Test set	144	119	80	518
Training set (%)	16.25	10.68	17.28	55.76
Test set (%)	16.72	13.82	9.29	60.16
All Data (%)	16.36	11.44	15.36	56.82
Related Only (%)	37.91	26.50	35.58	-

Table 6.4: Data distribution in the FNC-SE corpus

[9] `https://figshare.com/s/d87b42713a55e56c823a`

To choose an expertise value for a source, we chose the label selected by at least two annotators in the annotation. We observed that amongst 645 annotations, two or more annotators agreed on 466 items for the same expertise label. There are 179 items where none of the annotators' labels match. To select an appropriate label for the source expertise, I annotated 179 items in isolation. Thereafter, my source expertise label that matches with any one of the previously done annotations is selected as the final label for the source expertise. Although, I did the final labelling of the source expertise, my experiment is not affected by it. This is because ARs with *0: Not an expert* and *CD* labelled sources are not considered for the stance detection. I did this to study the role of expert sources in detecting stance.

6.7 Role of expert sources in the stance detection: a computational approach

To analyse the role of source expertise in the stance detection, first I analyse how the model works while using all data in FNC-SE corpus. Thereafter, I remove attribution relations that have sources whose expertise are *0*. Additionally, I also remove ARs that have sources expertise labelled as *CD* considering that indecisive state of participants might not contribute in the stance detection. I implemented the same system architecture that I used in Section 4.5 for stance detection.

I used Robustly Optimised BERT approach (RoBERTa) (Liu et al. 2019) for the stance detection task. I used Simple Transformers by Rajapakse (2017) that is built on top of Hugging Face transformers to use a pre-trained RoBERTa model. I used headlines and concatenated attributions as the input. I trained transformers for 10 epochs with a batch size of 4 and a learning rate of $1e^{-5}$. The output of the model is a stance label. I performed a two-stage classification. At the first stage, the stance is classified as $\mathrm{related}$ or $\mathrm{unrelated}$. At the second stage, I considered data classified as related from the first stage. The output of the second stage is any one of the stance labels agree, disagree or discuss. At both stages, I implemented the same machine learning model. Figure 4.3 for two-stage architecture is previously explained Section 4.5 of Chapter 4.

At the first stage of the stance detection, I observed that classification results for unrelated class is almost 100%. My approach for the first stage was able to reliably determine whether the headline and article are on the same topic. Thus, I can say that a stance detection system can reliably segregate the unrelated data from related ones. Now, at the second stage my objective is to classify related data as any one of three labels agree, disagree or discuss.

To assess the role of source expertise in stance detection, I hypothesised that,

> H6.1: Removal of ARs with non-expert sources helps the stance detection model to learn the correct context for the final prediction.

As per my hypothesis, the driving element for the stance in a news body could be claims coming from the source with higher expertise. Moreover, in such context, claims coming from low expertise sources may confuse the machine learning model to learn wrong samples and end up with poor results. To clarify, let us take an example.

> **Example 6.2.**
> **Headline**: Tropical spider burrowed under man's skin through appendix scar and lived there for THREE DAYS
>
> **News body extract with ARs**: Prepare to meet ... mite man. Doctors removed a matchhead-sized insect, believed to be a spider, from under Dylan Thomas's skin earlier this week and have sent the creature away for testing to determine what it is. It had been there for three days and burrowed up to his chest, leaving a trail of red blisters. The 21-year-old was on his first trip to Bali. **He** told News Corp yesterday *that doctors had pulled a tropical*

> *spider "a bit bigger than the size of a match head" from his skin.* There's just one problem. *Spiders,* according to **Perth arachnid expert Dr Volker Framenau***, don't burrow in skin. "They don't have the tools, the armature, to do this sort of stuff,"* **Dr Framenau** said. *"I find it highly unlikely, almost impossible, that it was a spider.' More likely,* **Dr Framenau** said*, was some kind of burrowing mite. "That's a professional skin-digger,"* **he** said. *"There's a lot of nasty stuff out there."* The results of the tests on the creature should come back next week. Mr Thomas has been contacted for comment.

The stance class for Example 6.2 is disagree in the FNC-SE corpus. We can see in the example that the stance of the news body to the headline is based on what an expert source **Perth arachnid expert Dr Volker Framenau** is claiming, but not on **the 21-year-old Dylan Thomas**'s claims. Despite **the 21-year-old Dylan Thomas** being the man from whose skin doctors pulled a creature and he told about the incident, the stance label is based on **Dr Framenau**'s claims of that creature not being a spider. With this observation, I can say that expert sources affect the stance labelling. Additionally, I can say that sources that are labelled as *not an expert* changes the stance direction. In Example 6.2, if we follow what **the 21-year-old Dylan Thomas** is claiming then the stance label should be agree. Non-expert sources are represented by the label *0: Not an expert* during annotation. Table 6.5 shows how often non-expert sources occurs at least once in different stance classes of training and test sets of the FNC-SE corpus. As per the Table 6.5, the training set has a quarter of data with at least an attributional source which is not an expert. The test set has even higher occurrence of such non-expert sources. More than half of test data has at least one non-expert source.

	Agree	**Disagree**	**Discuss**	**Total**	**FNC-SE data**	**%**
Training	97	95	108	300	1200	25
Test	86	68	42	196	343	57.14

Table 6.5: Data distribution in FNC-SE related stance classes with source expertise = *0: Not an expert*

As per Table 6.5, I can say that frequently occurring non-expert sources makes their claims less credible to readers and hence, make the credibility of the news document questionable. Additionally, such frequently occurring non-expert sources also make the stance decision complex for the computational model. As I hypothesised that the non-expert sources' claims could confuse the stance detection model by not letting it to learn the correct context, I decided to remove ARs that have sources labelled as *0: Not an expert*. In this way, we could make the machine learn from the correct samples and learn to capture the right context.

To analyse how the stance detection is affected by expert sources, first I trained the stance detection model using all attribution relations in the news body of each document. I tested the trained model on all test set data. Additionally, I also separately tested the same model only on those test data that have at least a non-expert source. In both cases, I got a macro average F-score of 0.58 as shown in Table 6.6. As per my hypothesis, I removed such ARs that have source expertise labelled as *0: Not an expert* from the training set of the FNC-SE corpus. Thereafter, I trained the stance detection model with the training data that only have expert sources. I anticipated that the model learning from the correct context may yield better test results. In Table 6.6, we can see that the model trained with ARs with only expert sources (excluding ARs with non-expert sources) performs better than the previous model that was trained with all data. The model using only expert sources has improved results for the test set by 14%. Additionally, I also checked how the model performs for those test data that have at least a non-expert attributional source. The model performance results are shown in Table 6.6. The model is successful in capturing the right context also in such data, increasing the macro average F-score by 18% over the previous

Model	**Test Data**	**Precision**	**Recall**	**F-score**
Trained using all ARs	All (343)	0.6	0.61	0.58
	Atleast one 0 (196)	0.58	0.56	0.58
Trained removing ARs with SE = 0	All (343)	0.73	0.72	0.72
	Atleast one 0 (196)	0.77	0.77	0.76

Table 6.6: Stance detection results in related data with/without using source expertise

model.

	Precision	**Recall**	**F-score**
agree	0.59	0.85	0.70
disagree	0.51	0.30	0.38
discuss	0.77	0.61	0.68

Table 6.7: Stance detection performance with all ARs

	Precision	**Recall**	**F-score**
agree	0.75	0.73	0.74
disagree	0.60	0.70	0.65
discuss	0.86	0.70	0.77

Table 6.8: Stance detection performance while removing ARs with Source Expertise = 0

6.8 Discussion and Error Analysis

Table 6.7 and Table 6.8 show that there is a considerable improvement on the disagree stance classification. The improvement happens after removing ARs from the training input that have source expertise labelled as *0: Not an expert*. As per my hypothesis, the model might have learn from training data that lacks ARs with non-expert sources. The stance detection model might have learned to capture the context for correct classification. To clarify, let us take an example from the training set of the FNCSE corpus.

Example 6.3.
Headline: 'Batmobile Stolen From "Batman v Superman: Dawn of Justice" Set, Zack Snyder Knows Who Did It

Stance: disagree

News Body with ARs:
On Friday, **bleedingcool.com** said, *"The scuttlebutt from sources in Detroit is that one of the Batmobile models being used in the filming of Batman Vs. Superman has gone missing, believed stolen."*
Not surprisingly, the Internet went into a tizzy, but later that day, **Detroit police** said *the theft was a rumour*.

Sgt. Michael Woody told the Detroit Free Press *that police confirmed with producers of Batman v. Superman: Dawn of Justice that the vehicle has not been stolen.*
"The Batmobile is safe in the Batcave where it belongs," **Woody** said.
The paper also said *that sources close to the movie being filmed in D-Town also said the fly ride had not been stolen.*

In the Example 6.3, all attributional sources have a certain level of expertise except the source **bleedingcool.com** which is annotated *0: Not an expert*. **bleedingcool.com** is the source that makes the context confusing as it talks about missing of a Batmobile model that was used in the filming of Batman Vs. Superman. In contrast rest of the news body is about the theft being a rumour and the vehicle was not stolen. Expert sources like **Sgt. Michael Woody** clarified that the vehicle is safe and theft of it is a rumour. The AR with a non-expert source **bleedingcool.com** if removed from the consideration let the model train from the right context.

I tested the stance detection model on the test set of the FNC-SE corpus but, without removing ARs with non-expert sources. Although I removed non-expert sources from training data, I did not do the same in the test set to see how the model performs in the presence of noise. As I anticipated, the model showed an encouraging performance as reported in Table 6.8. Removing ARs with non-expert sources from training data considerably increased the disagree stance classification. Following is an example that was classified as agree while the model was trained with all ARs but classified as disagree after removing ARs with non-expert sources from the training data.

Example 6.4.
Headline: Argentina's President Cristina Kirchner Adopts Jewish Godson To Prevent Him Turning Into A Werewolf

Stance: disagree

News Body with ARs:

> *There's an old Argentinian legend that a seventh child will turn into "el lobizon" — aka a werewolf — after his 13th birthday, and then terrorize the Argentinian countryside at night whenever there's a full moon*, as reported by **the Independent**.
>
> Over the last few days, **sources** have been reporting *that this Argentinian custom was adopted as a response to the murder and abandonment of these "el lobizon" children.*
>
> *Reportedly, the godchild custom goes back all the way to 1907 when Russian emigrés asked the then-president José Figueroa Alcorta to become the godfather to their seventh son,* reports **the Guardian**.
>
> *"The local myth of the lobizón is not in any way connected to the custom that began over 100 years ago by which every seventh son (or seventh daughter) born in Argentina becomes godchild to the president,"* **Argentine historian Daniel Balmaceda** told The Guardian.
>
> *Fernández has become the president godmother to roughly 700 children since she took office in 2007,* reports **The Guardian**.

In Example 6.4 **sources** is the one that is annotated as *0: Not an expert*. The rest of the sources are annotated with a certain level of expertise (1 to 4). However, during classification all ARs are provided in the input. The model trained excluding ARs with non-expertise sources seems to capture the right context for correct classification.

Additionally, I observed that headlines with preventive words (like stop, prevent) and negation (like not, no), that were found problematic in stance classification(see Section 4.7 for details), are detected with the correct stance labels. For instance, the news body in Example 6.4 is classified correctly with headlines like *Argentina's President Just Adopted a Son So He Won't Turn into a Werewolf, 'Argentina's President Adopted A Jewish Godson To Stop Him From Turning Into A Werewolf, Argentina's President Adopts Young Boy so He Won't Turn Into Werewolf.* Such results show that my stance detection model trained using ARs with expert sources can handle some errors identified in the stance detection model trained using all ARs (as in Chapter 4).

6.9 Conclusion

I found that attribution relations with expert sources are useful in stance detection. In this chapter, I enhanced a subset of the FNC-1 corpus with source expertise data using Amazon Mechanical Turk. As I used a subset of the FNC-1 corpus for this task that has only stance annotations, I annotated attribution relations in the subset corpus FNC-SE using the guidelines by Newell, Margolin & Ruths (2018). To validate my annotation, one of my supervisors annotated around 10% of randomly selected data from the corpus using the same annotation guidelines. In the attribution annotation task, agreement coefficient with 0.865 validates the reliability of my annotation.

I conducted 23 surveys that have four news documents in each survey. Each survey is annotated by at least three annotators who provided labels for the source's level of expertise. I observed a moderate agreement of 0.519 Gwet AC2 for source expertise annotation in the FNC-SE corpus.

The objective for creating the FNC-SE corpus was to analyse the role of source expertise in the stance detection. Thus, I performed experiments to analyse how the stance detection model performs while trained with and without non-expert sources. My results show that the presence of non-expert sources and their claims might mislead the model and result in wrong predictions. Training the model with expert sources might be helpful for learning to capture right context for the stance detection. In conclusion, I can say that source expertise can be useful to train the model with correct context and hence, it is has an important role in the stance detection.

The contribution of this chapter is two-fold:

i. I prepared an extended subset dataset of FNC-1 corpus and enriched that with source expertise data.

ii. I showed that source expertise information is useful in stance detection.

Chapter 7

The role of attributional cues in stance detection

7.1 Introduction

As described in Section 3.1, an attribution relation (AR) is composed of three key components- **source**, cue and *content*. In this chapter, I focus on the component cue and analyse its contribution in the stance detection. A cue is a lexical content that connects a **source** to its attributed *content*. Cue expresses a **source**'s attitude that may be an assertion, knowledge, belief or intention. It means a cue expresses the stance of the **source** towards its attributed *content*. Take the following example to see effects of cues.

> **Example 7.1.**
> **KLAS-TV** in Las Vegas is reporting that *Jose Canseco was accidentally shot in his home*. **A neighbor** tells 8 News NOW *former baseball star Jose Canseco was hurt in an accidental shooting Tuesday afternoon at his house on the eastside of the Las Vegas valley*. **Metro Police** confirm *there was an accidental shooting at the address*, but would not confirm *the former player was hurt*.

In Example 7.1, there are two sources: **KLAS-TV** and **A neighbor** who are claiming that a former baseball player was accidentally hurt in his home.

However, their stance towards their claims is neutral. Moreover, their commitment to their claims is not strong. In contrast to their claims, a third source **Metro Police** was certain that there was an accidental shooting but, he is doubtful that the player was hurt. We can see that **Metro Police**'s commitment to his claims is strong because of the use of cues like confirm. The appearance of such commitment cue with an expert source for that context like **Metro Police** allows news readers to judge the trustworthiness of the claim. Thus, commitment in the cue can be useful to decide how much weight a reader should give to the *claim* while reading a news article. On this basis, I decided to analyse the role of such cues in stance detection.

The research questions that I intend to explore in this chapter is as follows.

> RQ5: Are attributional cues useful in the stance detection?

To explore this research question, firstly I hypothesised that,

> H5.1: The stance detection performance of the machine learning model might degrade when cues are removed from the input.

To test this hypothesis, I compared the stance detection performances of the machine learning models with and without using cues in the input. In this way, I analysed the role of cue in stance detection by assessing how the absence of cue in an attribution relation can affect the stance detection model performance.

Additionally, I analysed how frequent Ghanem et al. (2018)'s cues occur as cues in attribution relations. Ghanem et al. (2018) showed that their cue list is useful in the stance detection. In this work, I analysed how often attributional cues with certainty, doubt or neutrality occur in three stance classes: agree, disagree and discuss of the FNC-1 corpus. I define certainty, doubt and neutrality cues as follows:

- certainty: cue that shows certainty or confidence of source towards their claims or attributed contents. For example, *confirm* and *clarify*.
- doubt: cue that expresses doubt or scepticism of source towards their claims or attributed contents. For example, *speculate* and *think*.

- neutrality: cue that is used for reporting and source's feel of neutrality towards their claims or attributed contents. For example, *say* and *tell*.

For the certainty, doubt and neutrality cues, I hypothesised followings:

> H5.2: doubt cues appear more often in the disagree stance class.
>
> H5.3: certainty cues are more often associated with agree stance class.
>
> H5.4: neutrality cues are the most frequent cues in all stance classes with its higher ratio contribution in discuss class.

7.2 Literature Review

Previous work on stance detection (Bahuleyan & Vechtomova 2017, Ghanem et al. 2018) showed that hand-curated cue features are useful to classify stance in short-texts (like tweets) as well as in long-texts (like news documents). I used cues originally introduced by Bahuleyan & Vechtomova (2017) to create cue categories in my work. Bahuleyan & Vechtomova (2017) prepared 153 hand-curated cue features under 9 different categories that are belief, report, doubt, knowledge, denial, curse words and internet slangs, negation words and questions words. They collected cues by manually inspecting the training set of Subtask-A: SDQC for RumourEval, task- 8 of SemEval 2017. They argued that the presence of such cues in tweet replies could indicate their type that is whether a reply tweet is supporting, denying, commenting or querying the underlying rumour. It means cue features give the stance of tweet users towards their tweets.

Ghanem et al. (2018) used six cue categories: *belief*, *report*, *doubt*, *knowledge*, *denial* and *negation* from the cue list by Bahuleyan & Vechtomova (2017). Ghanem et al. (2018)'s work added a seventh cue category Fake that has a combination of some words from the FNC-1 baseline polarized words list and words from the cue list by Bahuleyan & Vechtomova (2017). Ghanem et al. (2018)'s work showed that adding cue features increases stance classification performance in the FNC-1 corpus[1]. Ghanem et al. (2018) extracted cue features from news articles, excluding headlines. Ghanem et al. (2018) computed information gain to show how important each cue category is for each stance class. They found that there is not any meaningful importance order of cue categories for the unrelated class. They argued it could be because the unrelated class contains randomly paired headlines and articles that belong to different topics. In the case of related headline-article pairs, there is a clear significance order of cue features. For instance, the disagree class, cue categories *fake* and *denial* appear to be more important than other cues. Similarly, the agree class article bodies are dominated by *report* and *belief* cue features. As Ghanem et al. (2018)'s cue categories are found useful in stance detection, I decided to analyse how frequent those cues appear as attributional cues in the same news documents.

Pareti (2015) represented cues of attribution relations as a way to express authorial stance. Pareti (2015) defined authorial stance as the author's commitment toward the truth of information expressed in contents. If the contents are considered as truthful, the authorial stance is labelled as *committed*, if not then *not committed* or as default *neutral*. Examples of those labels from Pareti (2015)'s work are as follows.

> Committed: admit, know
>
> **Mr Abbott** knew *that Gillard was in Sydney*.
>
> Not committed: lie, joke
>
> **Dr. Smith** jokes: *"There is no correlation between smoking cigarettes and the incidence of lung cancer in the population."*
>
> Neutral: say, believe
>
> **A Boeing spokeswoman** said *a delivery date for the planes is still being worked out "for a variety of reasons, but not because of the strike."*

Soni et al. (2014) argued that cue words signal the factuality of claims. Soni

[1] `https://github.com/FakeNewsChallenge/fnc-1`

et al. (2014) used cues (or predicates) associated with different sources to assess the factuality of associated claims. From the literature review, we can say that cues are a useful bit of information in a news article. I therefore analysed whether the removal of attributional cues in a news article has any effects on the performance of a stance detection system. I discuss this in the next section.

7.3 Analysing the role of attributional cues

Cues in an attribution relation express the propositional attitude of the source towards what being claimed (Pareti 2015). In short, an attributional cue expresses the source's commitment to their claims. In this section, we are going to analyse the role of attributional cues in the stance detection of news articles by comparing the model performance with and without cues in the input.

I analysed the stance detection model's performance with and without using cues in attribution relations. The objective of this experiment is to see if the system performance degrades while removing cues from attribution relations. Here, for the stance detection, I implemented the same two-stage system architecture with transformer-based models discussed in Section 4.5.

		Predicted			
		agree	disagree	discuss	unrelated
Actual	agree	1267	61	480	50
	disagree	201	297	155	16
	discuss	516	108	3647	102
	unrelated	68	8	170	17802

Table 7.1: Confusion matrix while using Attribution Relations without cues as input

The overall labelling done by the two-stage stance detection system is represented as a confusion matrix in Table 7.1. Here I used only sources and claims of attribution relations with the news headlines as the input to the stance detection system. I excluded cue texts from attribution relations.

		Predicted			
		agree	disagree	discuss	unrelated
Actual	agree	1254	67	491	46
	disagree	150	311	195	13
	discuss	331	103	3878	64
	unrelated	58	5	124	17861

Table 7.2: Confusion matrix while using Attribution Relations with cues as input

Table 7.3 reports the performance of the model for each stance class labelling. The results are biased to the data distribution in the FNC-1 corpus with the best performance for the unrelated class. For a clear comparison, Table 7.2 and Table 7.4 show the confusion matrix and performance of the stance detection model including cues in the input.

	Precision	**Recall**	**F-score**
agree	0.62	0.68	0.65
disagree	0.63	0.44	0.52
discuss	0.82	0.83	0.83
unrelated	0.99	0.99	0.99

Table 7.3: A two-stage stance detection model performance using Attribution Relations without cues as input

	Precision	**Recall**	**F-score**
agree	0.70	0.67	0.69
disagree	0.64	0.46	0.54
discuss	0.83	0.89	0.86
unrelated	0.99	0.99	0.99

Table 7.4: A two-stage stance detection model performance using Attribution Relations with cues as input

Now, to analyse how usable are cues in stance detection, I compared my experiment results of stance detection with and without using attributional cues in the input. Table 7.5 shows the comparison which shows that the overall performance of the stance detection model decreases by 3% while excluding cues from the input attribution relations. It illustrates that attributional cues are a useful bit of information in detecting the stance of a news article to its headline.

	class-wise F-score				
	agree	**disagree**	**discuss**	**unrelated**	**macro F-score**
ARs with cues	0.69	0.54	0.86	0.99	0.77
ARs without cues	0.65	0.52	0.83	0.99	0.74

Table 7.5: Comparing the stance detection performance of the model with and without attributional cues

7.4 Attributional cues in the FNC-1 corpus

Before analysing how commitment expressed in attributional cues (represented as cue throughout the dissertation) affects the stance prediction, I analysed the distribution of different cues in the FNC-1 corpus. For that, first I applied AR model discussed in Chapter 3 to the FNC-1 corpus to detect all attribution relations. In this section, I discuss the distribution of those AR constituents that are tagged as cue by the AR model.

The AR model detected 138,165 cues in the training set and 65,623 cues in the test set of the FNC-1 corpus. To find the top 30 most frequently occurring cues in the Training set, I used the Spacy Lemmatizer[2] to convert words to their root forms. The cue list along with the number of their appearance is given in Appendix D.1. Similarly, the top 30 most frequently appearing cues in the test set are presented in Appendix D.2.

Table 7.6 shows number of cues in the training and test sets of the FNC-1 corpus expressed per the three related stance classes. I didn't include details for the unrelated class because the stance detection model is almost 100% effective to correctly classify the unrelated class data. Additionally, Ghanem et al. (2018) highlighted that it is not possible to describe which cues are important in the unrelated class because the unrelated class has topically different headlines and bodies paired to form a news document. I therefore decided to analyse the role of cue in classes that have related headline-body pairs.

Pomerleau & Rao (2017) build the FNC-1 corpus by associating the same news bodies with different headlines, resulting in repetition of cues that

[2] `https://spacy.io/api/lemmatizer`

	Training Set	**Test Set**
agree	31,538	18,301
disagree	7,808	7,671
discuss	98,819	39,651

Table 7.6: The cue distribution in the FNC-1 corpus as per stance classes

are detected by the AR model. There are respectively 4,737 and 3,070 different and unique cues in training and test sets of the FNC-1 corpus. The distribution of those unique cue per stance class is given in Table 7.7.

	Training Set	**Test Set**
agree	2,481	1,890
disagree	1,050	1,138
discuss	3,971	2,220

Table 7.7: Unique cue distribution in the FNC-1 corpus as per stance classes

The top 20 most frequently appearing cues in the agree, disagree and discuss classes in the training and test sets of Table 7.7 are given in Appendix D.3 and Appendix D.4 respectively. The most frequent cue is at the top of the table and frequency decreases while moving down towards the bottom of tables.

To analyse which cues are unique and exclusive per stance class, I extracted cues that appear in a stance class but, not on rest of the classes. Table 7.8 reports numbers of cues that are unique and exclusive per stance class. Table 7.8 shows that there is a small number of cues which are exclusive to the disagree class. This implies that detecting the disagree class data from rest of the classes is a difficult task because ARs in disagree class news bodies share cues with other stance classes.

	Training Set	**Test Set**
agree	508	348
disagree	16	6
discuss	2,214	1,157

Table 7.8: Number of cues unique and exclusive per stance class in the FNC-1 corpus

7.5 Cue occurrence as an attributional cue in the FNC-1 corpus

In this section, we explore whether Ghanem et al. (2018)'s cue list is effective for differentiating stance classes as per their occurrence as cue. Ghanem et al. (2018) extended cue features implemented by Bahuleyan & Vechtomova (2017) using word embeddding and analysed contribution of each cue category in stance classification task. Bahuleyan & Vechtomova (2017) collected 153 cues from Twitter data and formed nine categories of those cues. Using Bahuleyan & Vechtomova (2017)'s cues with an additional category *fake*, Ghanem et al. (2018) analysed contribution of seven cue categories to classify stance of news articles as unrelated, agree, disagree and discuss. In their work, cue features are extended by adding five most relevant words extracted using word2vec. Ghanem et al. (2018)'s cue categories along with total number of cues before and after pre-processing are shown in Table 7.9. During pre-processing, I converted all cues to lowercase and changed them to their root word using the Spacy lemmatizer [3]. I did the pre-processing because I observed that there are many cues with the same root words but, are used as different cues such as *believing*, *believe*, *believed*. I also observed that the same cue with first letter uppercase and lowercase are treated as two different cues such as *Evidence*, *evidences*, *Possibly*, *possibly*.

	Total cues	Pre-processed cue
Belief	107	69
Denial	117	60
Doubt	132	83
Report	183	95
Knowledge	85	55
Negation	512	121
Fake	158	88

Table 7.9: Different categorical cues in extended lists by Ghanem et al. (2018)

Table 7.10 shows example cue features belonging to seven cue categories in Ghanem et al. (2018)'s cue list. In extended cue features by Ghanem

[3] `https://spacy.io/api/lemmatizer`

	Example cue features
Belief	believe, think, consider
Denial	refuse, reject, dismiss
Doubt	unsure, guess, doubt
Report	say, assert, claim
Knowledge	confirm, support, admit
Negation	nothing, never, don't
Fake	false, rumour, hoax

Table 7.10: Cue feature examples (Ghanem et al. 2018)

et al. (2018), I observed that some cue categories share the same cues. For example, cue categories *belief* and *doubt* share same 17 cues that are as follows.

> possibility, presume, suppose, possible, ponder, say, know, probably, assume, anyway, maybe, believe, potential, guess, infer, think, surmise

The number of common cues is more than 20% of cues in each category *belief* and *doubt*. The repetition of the same cue features in different categories could make it difficult to analyse how important is a cue to a stance class. Number of cues common among different categories are given in Table 7.11.

	Denial	**Doubt**	**Report**	**Knowledge**	**Negation**	**Fake**
Belief	0	17	9	6	3	1
Denial		2	1	4	0	8
Doubt			13	7	2	4
Report				13	0	2
Knowledge					2	4
Negation						10

Table 7.11: Cues common among different categories

Despite several cues being shared amongst different categories, I analysed how often Ghanem et al. (2018)'s cues occur as <u>cue</u> (attributional cue) in three stance classes of FNC-1 corpus. For the comparison, I used pre-processed cue features (lemmatised and lower-cased). Table 7.12 shows frequency of seven cue categories in different stance classes of the FNC-1 corpus training set.

	agree	disagree	discuss
Belief	7447	1987	25191
Denial	138	29	507
Doubt	9660	2613	31792
Report	11976	3095	39169
Knowledge	7840	2098	25898
Negation	238	70	779
Fake	157	41	380

Table 7.12: Ghanem et al. (2018)'s cue occurrence in the FNC-1 corpus stance classes

Figure 7.1, Figure 7.2 and Figure 7.3 illustrate distribution of Ghanem et al. (2018)'s cue categories as cue (attributional cue) in agree, disagree and discuss stance classes in the training set of the FNC-1 corpus. The figures show that four cue categories *belief*, *doubt*, *report* and *knowledge* appear most frequently in all three stance classes. I did not observe any considerable difference in the distribution of Ghanem et al. (2018)'s cues as cues in those three classes.

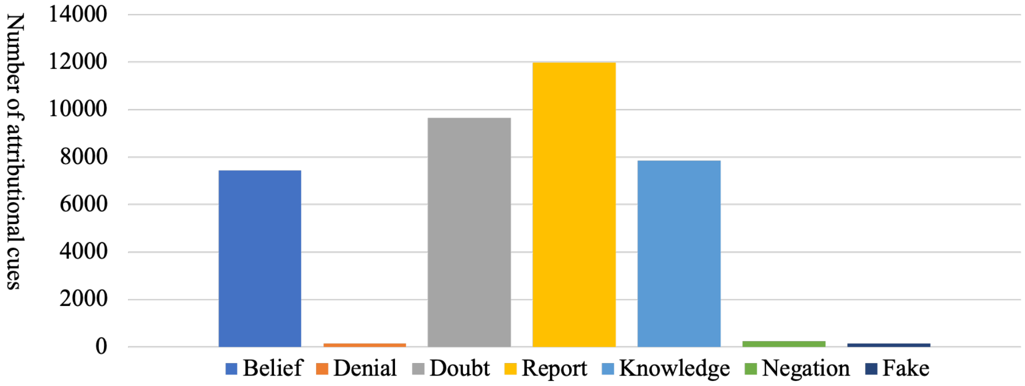


Figure 7.1: Cue categories distribution in agree stance class

	agree	disagree	discuss
Belief	19.88	20.00	20.36
Denial	0.36	0.29	0.40
Doubt	25.79	26.30	25.69
Report	31.97	31.15	31.66
Knowledge	20.93	21.12	20.93
Negation	0.63	0.70	0.62
Fake	0.41	0.41	0.30

Table 7.13: Occurrence % of Ghanem et al. (2018)'s cue occurrence in the training set of FNC-1 corpus

The percentage (%) distribution of attributional cues per Ghanem et al.

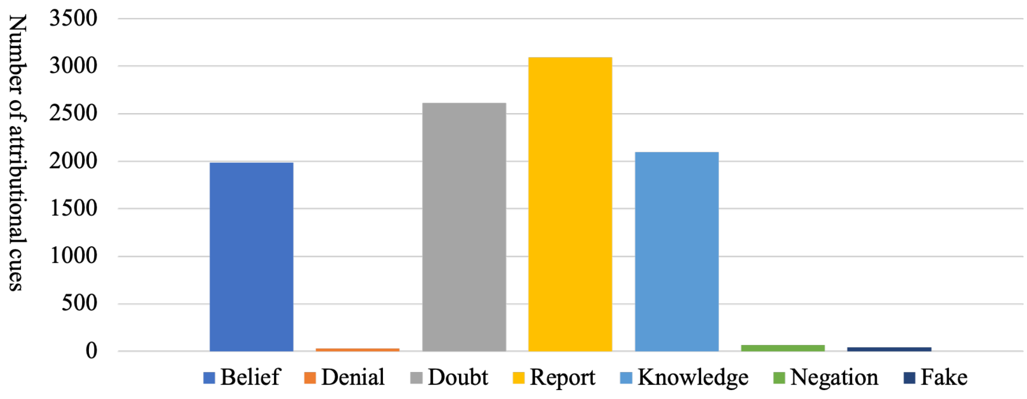


Figure 7.2: Cue categories distribution in disagree stance class

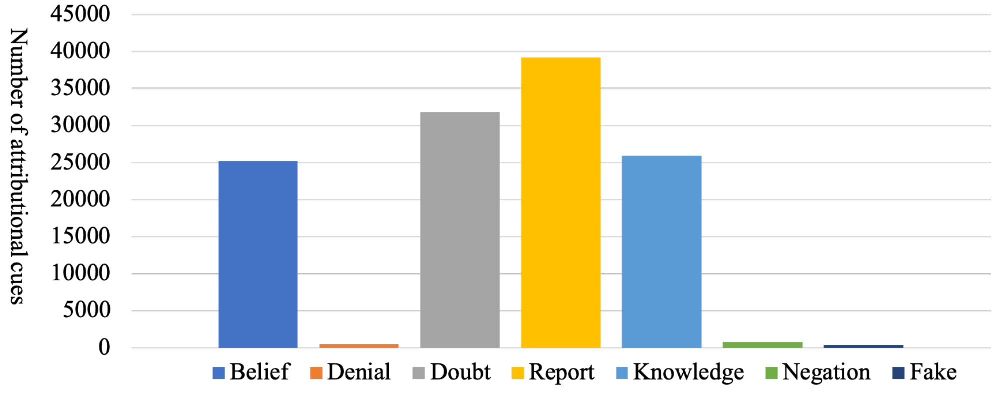


Figure 7.3: Cue categories distribution in discuss stance class

(2018)'s cue categories in the three stance classes agree, disagree and discuss classes of the FNC-1 training set is given in Table 7.13. Bahuleyan & Vechtomova (2017) argued that presence of *belief* or *knowledge* words could indicate a tweet reply where their authors express support. Similarly, *doubt* or *denial* word cues are used to show disagreement (Bahuleyan & Vechtomova 2017). Following this literature, I anticipated that *belief* and *knowledge* cues showing certainty are more likely to appear in the agree class. Similarly, doubt expressing cue categories like *doubt* and *denial* occur more often in the disagree class. Furthermore, discuss class might have neutrality expressing cues from *report* category. However, Table 7.13 shows that all cue categories are equally likely to occur in all three stance classes, but with a small skew of *doubt* cue percentage for disagree class.

Ghanem et al. (2018) used a list of 1294 cues divided into seven different categories and showed that those cue categories have different order of

	Total unique cues	**Missed unique Cues**	**Total Cues**	**Missed total cues**	**Missed %**
Training	4745	1452	138165	28520	20.6
Test	3070	962	65623	13969	21.2

Table 7.14: Ghanem et al. (2018) Cue occurrence as attributional cues in Training and Test sets of the FNC corpus

importance for each each stance class in the FNC-1 corpus. I analysed how often cues by Ghanem et al. (2018) occur as cues in training and test sets of the FNC-1 corpus whose summary is given in Table 7.14. As per the Table 7.14, more than 20% of attributional cues in both training and test sets of the FNC-1 corpus are not captured by the Ghanem et al. (2018)'s cue list. This could be because cues in Ghanem et al. (2018)'s list were originally collected from twitter data from a dataset annotated for stance detection and veracity prediction of rumours.[4] Additionally, all Ghanem et al. (2018)'s cues are not attributional cues.

7.6 Conclusion

Removing cues from the input attribution relations decreases the overall performance of stance detection model by 3% F-score. This shows that attributional cues are useful additional information to detect the stance of a news body towards its headline. Additionally I observed that Ghanem et al. (2018)'s cue list does not seem to be effective to differentiate stance classes as per their occurrence as cue. Moreover, I did not observe any significant association of certainty, doubt and neutrality cues with agree, disagree and discuss classes respectively. However, doubt expressing cues showed a small skew in the disagree class.

[4] https://alt.qcri.org/semeval2017/task8/

Chapter 8

Conclusion

In this chapter, I discuss the conclusions for my dissertation. First, I revisit my research questions and summarise the findings. Second, I consider the limitations of my work. Finally, I explore the future directions for my research.

8.1 Research Questions and Findings

RQ1: How can we detect attribution relations in a news document?

Findings: As discussed in Chapter 3, we can detect each component- **source**, cue, *content* of attribution relations in a news body by using a new model for attribution relation detection. The model is a sequence tagging system that implemented ELMO embedding, two bi-directional long short-term memory networks and a densely connected neural network to detect **source**, cue and *content* in news bodies. The model's performance outperformed the baseline by an accuracy of 36% for token-wise and 43% for sentence-wise predictions. Additionally, a test for the broader applicability of the AR model showed that the model is effective also on a different domain, specifically the Vaccination corpus. The Vaccination corpus not only contains news documents but, also editorials, blog content etc. However, the attribution relation detection performance on the Vaccination corpus is 8.45% and 12.33% less accurate respectively for token-wise and sentence-wise predictions than on the test set of the PolNeAR corpus. The AR model is trained using the training samples from the PolNeAR corpus. Such findings imply that the AR model is applicable in the FNC-1 corpus to extract attribution relations that have only stance annotations.

RQ2: Are attribution relations useful to detect stance of a news body to its headline?
Findings: As discussed in Chapter 4, using attribution relations, instead of a whole news body as input to the stance detection system showed comparable performance to the current best systems. Here, the stance detection system is a pre-trained transformer-based RoBERTa-large model. My stance detection model's performance is 6% less than that of the state-of-the-art (Zhang et al. 2019). However, my system's performance for the agree and discuss classes outperformed Zhang et al. (2019)'s results by the macro-F of 2% and 3% respectively with the same performance for unrelated class. Zhang et al. (2019) proposed a stance detection system that is focused on correctly classifying the minority class disagree. Additionally, I found that attributed *content*s (also known as claims in my work) alone are effective in stance detection that lessen by 3% macro-F than while using whole ARs. The uneven distribution of stance classes in the FNC-1 corpus has made the classification problem complex. The most frequently appearing stance class unrelated covers 73% of the FNC-1 corpus. In contrast, the least appearing disagree class contains less than 2% of total data in the FNC-1 corpus. The comparison of my experimental results with state-of-the-art results for unrelated and discuss classes show that class-specific features can be helpful to get good results for minority classes agree and disagree. Additionally, my error analysis showed that the presence of the negating words (like *not*, *no*) and preventive words (like *stop*, *prevent*) can make a piece of text disagree with another are difficult to capture.

RQ3: Is a reader's judgement of claim credibility correlated with his/her judgement of the level of expertise of the source who is making that claim?

Findings: Yes, an empirical study showed that there exists a correlation between the claim credibility judgement and the source expertise judgement. As described in Chapter 5, I statistically tested whether people's judgements of claim credibility is correlated to their judgement of the source's level of expertise.

RQ4: What is the role of source expertise information in detecting the stance of a news body to its headline?
Findings: Source expertise information is useful in the stance detection. I observed that removing attribution relations with non-expert sources from the training data helps the stance detection model learn the right context. This is reflected in the results with an increment of 14% macro F in the test set and an increment 18% macro F in only those test data that have at least one non-expert source. Such experiment results (details in Chapter 6) showed that expert sources are the driving element for the stance prediction. Thus, I argue that attribution relations with expert sources contribute to correctly classifying the stance of a news body to its headline. The training and test set data in experiments belong to the FNC-SE corpus. The FNC-SE corpus is a subset of the FNC-1 corpus Pomerleau & Rao (2017) that is enriched with source expertise data collected through crowd sourcing.

RQ5: Are attributional cues useful in the stance detection?
Findings: Yes, cues are a useful piece of information as experimental results show that their removal from attribution relations in the input degrades the performance of the stance classification by 3% macro F. In addition, I observed that Ghanem et al. (2018)'s cue words do not always appear as cue. My analysis showed that Ghanem et al. (2018)'s cue list captured less than 80% of cues in the FNC-1 corpus.

8.2 Limitations

In this section, I discuss the limitations that I feel had the greatest impact on my findings. There are primarily two limitations in this work:

i. The unavailability of comparative systems for the AR model evaluation.

ii. The unavailability of enough representative samples for each stance class.

I believe the above-mentioned first point (i.) is rather the state of the research in this area than a limitation. Within this area, there is development of several works (Pareti 2015, Newell, Cowlishaw & Man 2018) that are not openly available. My work openly provides data and codes for other people so that they can have insights and can explore more in the area. The unavailability of other works restricted me to get a comparative analysis of the AR model with other best systems. Despite of this limitation, I validated the effectiveness of the AR model on Vaccination corpus (Morante et al. 2020).

My research is also affected by the lack of a reliable and big dataset with equal representative samples for all stance classes. Although neural networks are very powerful and have state-of-the-art results in many natural processing tasks, for good a performance, machine learning systems need training sets which are representative of the population being analysed. As discussed in Section 4.2, the FNC-1 corpus is not a representative of the population because it is highly unbalanced with 73% data in the highest occurring unrelated class and less than 2% data in the least occurring disagree class. As the stance detection model is not trained with equal representative samples from all classes, the classification got biased to the highly occurring stance class unrelated. Furthermore, the model did not handle minority classes like agree and disagree properly due to the lack of enough representative samples in the training set. A contrasting feature of the unrelated class is that it is not realistic because the class is formed by associating topically different headline-body pairs. Thus, I believe that the unrelated class is not significant in the stance detection because published news barely contains the topically different headline and news body. However, the majority of data in the FNC-1 corpus belong to the unrelated class.

8.3 Future work

There are several future work directions for my research in this dissertation. Most importantly I believe that attribution relations (ARs) can be useful for fact-checkers and general readers of the news. Attribution relations contain claims, their sources and the source's commitment to the claim. Extracting ARs only leave behind background details and the author's own opinions in the news document. I believe ARs contain all important elements of a news body. Let us consider the news article in Example 1.1 of Chapter 1,

> **Example 8.1.**
> Prepare to meet ... mite man. Doctors removed a matchhead-sized insect, believed to be a spider, from under Dylan Thomas's skin earlier this week and have sent the creature away for testing to determine what it is. It had been there for three days and burrowed up to his chest, leaving a trail of red blisters. The 21-year-old was on his first trip to Bali. **He** told News Corp yesterday *that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin*. There's just one problem. *Spiders,* according to **Perth arachnid expert Dr Volker Framenau,** *don't burrow in skin. "They don't have the tools, the armature, to do this sort of stuff,"* **Dr Framenau** said. *"I find it highly unlikely, almost impossible, that it was a spider.' More likely,* **Dr Framenau** said *was some kind of burrowing mite. "That's a professional skin-digger,"* **he** said. *"There's a lot of nasty stuff out there."* The results of the tests on the creature should come back next week. Mr Thomas has been contacted for comment.

Despite of reading the whole news body, attribution relations with distinct **source**, cue and *content* could be useful for fact-checkers and general readers of news. These ARs show what different claims are made by whom in the news body and commitment shown by **source**s towards their claims. Additionally, readers and fact-checkers could also consider expertise of sources to decide if the source's claim is true or false. Thus, automatically detected ARs could provide essential and manageable information to guide readers and fact-checkers to get an insight and take an informed judgement about the news. To test that hypothesis, we can do a study with participants (fact-checkers and general readers) asked to judge whether automatically detected ARs from a news body are helpful to get a instant insight into the news and to make any further judgements. Further judgements could be useful to take actions like sharing news on social media, and assessing truthfulness of the news contents.

An interesting language pattern worth further exploration is the nested ARs. The nesting happens When a **source** makes a claim referring to another source. For example,

> **Example 8.2.**
> **Bloomberg** reports *that the Republican nominee has either won or tied among the group of voters making $ 100,000 or more*, according to **the Roper Center for Public Opinion Research**.

Here, **the Roper Center for Public Opinion Research** is a **source** that is claiming *that the Republican nominee has either won or tied among the group of voters making $ 100,000 or more* which is actually reported by another **source**-**Bloomberg**. I believe that the credibility of the claim is affected by the expertise of the actual source (here, **Bloomberg**) than the other **source** who is referring to the the same claim. Thus, it will be interesting to see how such language patterns are perceived by people and how they computationally affect the stance classification.

Finally, with the continuous development of new deep learning models, there is always a space to improve the performance of the AR model and the stance detection system. Another future work includes dealing with the minority classes of the FNC-1 corpus separately such that agree and disagree classes are trained with correct context samples. We can do it by implementing a three-stage model such that we can filter unrelated class from the rest related classes in the first phase. Thereafter, we can separate the second highly occurring discuss class from the rest minority classes with binary classification. In the last phase, we can implement class-specific features for agree versus disagree stance classification because those classes can be contrasted with their specific body contents.

Bibliography

Abhijnan Chakraborty, Bhargavi Paranjape, S. K. (2016), Stop clickbait: Detecting and preventing clickbaits in online news media, *in* '2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', pp. 9–16.

Aker, A., Derczynski, L. & Bontcheva, K. (2017), 'Simple open stance classification for rumour analysis', *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP* pp. 31–39.

Alkhalifa, R., Kochkina, E. & Zubiaga, A. (2021), Opinions are made to be changed: Temporally adaptive stance classification, *in* 'Proceedings of the 2021 workshop on open challenges in online social networks', pp. 27–32.

Amidei, J. (2021), Evaluating the Evaluators: Subjective Bias and Consistency in Human Evaluation of Natural Language Generation, PhD thesis, The Open University.

Anastasiou, L. & De Liddo, A. (2021), Making sense of online discussions: Can automated reports help?, *in* 'Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–7.

Andreas Hanselowski, Avinesh PVS, B. S. (2017), 'Fake News Challenge - Team Athene', `https://github.com/hanselowski/athene_system`.

Artstein, R. & Poesio, M. (2008), 'Inter-coder agreement for computational linguistics', *Computational Linguistics* **34**(4), 555–596.

Bahuleyan, H. & Vechtomova, O. (2017), Uwaterloo at semeval-2017 task 8: Detecting stance towards rumours with topic independent features, *in*

'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', pp. 461–464.

Benjamin Riedel, Isabelle Augenstein, G. S. (2017), 'Fake News Challenge - Team UCL Machine Reading', `https://github.com/uclnlp/fakenewschallenge`.

Bourgonje, P., Schneider, J. M. & Rehm, G. (2017), From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles, *in* 'Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism', pp. 84–89.

Canini, K. R., Suh, B. & Pirolli, P. L. (2011), Finding credible information sources in social networks based on content and social structure, *in* '2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing', IEEE, pp. 1–8.

Carlson, L. & Marcu, D. (2001), 'Discourse tagging reference manual', *ISI Technical Report ISI-TR-545* **54**, 56.

Chesney, S., Liakata, M., Poesio, M. & Purver, M. (2017), Incongruent headlines: Yet another way to mislead your readers, *in* 'Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism', pp. 56–61.

Conforti, C., Pilehvar, M. T. & Collier, N. (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)', pp. 40–49.

Conroy, N. J., Rubin, V. L. & Chen, Y. (2015), Automatic deception detection: Methods for finding fake news, *in* 'Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community', American Society for Information Science, p. 82.

DeMers, J. (2016), '59 Percent Of You Will Share This Article Without Even Reading It', `https:`

`//www.forbes.com/sites/jaysondemers/2016/08/08/59-percent-of-you-will-share-this-article-without-even-reading-it/#58f476562a64`.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G. & Zubiaga, A. (2017), SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, *in* 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 69–76.

Dor, D. (2003), 'On newspaper headlines as relevance optimizers', *Journal of pragmatics* **35**(5), 695–721.

Dungs, S., Aker, A., Fuhr, N. & Bontcheva, K. (2018), Can rumour stance alone predict veracity?, *in* 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3360–3370.

Elson, D. K. & McKeown, K. R. (2010), Automatic attribution of quoted speech in literary narrative, *in* 'Twenty-Fourth AAAI Conference on Artificial Intelligence'.

European Commission, D.-G. f. C. N. (2018), 'Fake news and disinformation online', *Flash Eurobarometer 464* .

Ferreira, W. & Vlachos, A. (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', pp. 1163–1168.

Flanagin, A. J. & Metzger, M. J. (2000), 'Perceptions of internet information credibility', *Journalism & Mass Communication Quarterly* **77**(3), 515–540.

Forum, W. E. (2014), 'Outlook on the global agenda 2014', `https://reports.weforum.org/outlook-14/top-ten-trends-category-page/`.

Gabielkov, M., Ramachandran, A., Chaintreau, A. & Legout, A. (2016), 'Social clicks: What and who gets read on twitter?', *ACM SIGMETRICS Performance Evaluation Review* **44**(1), 179–192.

Ghanem, B., Rosso, P. & Rangel, F. (2018), Stance detection in fake news a combined feature representation, *in* 'Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)', pp. 66–71.

Giffin, K. (1967), 'The contribution of studies of source credibility to a theory of interpersonal trust in the communication process.', *Psychological bulletin* **68**(2), 104.

Goodhue, D. L. & Loiacono, E. T. (2002), Randomizing survey question order vs. grouping questions by construct: An empirical test of the impact on apparent reliabilities and links to related constructs, *in* 'Proceedings of the 35th Annual Hawaii International Conference on System Sciences', IEEE, pp. 3456–3465.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K. & Derczynski, L. (2019), SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, *in* 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 845–854.

Grčar, M., Cherepnalkoski, D., Mozetič, I. & Kralj Novak, P. (2017), 'Stance and influence of twitter users regarding the brexit referendum', *Computational social networks* **4**, 1–25.

Gupta, A. & Kumaraguru, P. (2012), Credibility ranking of tweets during high impact events, *in* 'Proceedings of the 1st workshop on privacy and security in online social media', pp. 2–8.

Gwet, K. L. (2014), *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*, Advanced Analytics, LLC.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M. & Gurevych, I. (2018), 'A retrospective analysis of the fake news challenge stance detection task', *In Proceedings of the 27th International Conference on Computational Linguistics, pages 1859–1874, 2018* .

He, H., Barbosa, D. & Kondrak, G. (2013), Identification of speakers in novels, *in* 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Vol. 1, pp. 1312–1320.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Hovland, C. I., Janis, I. L. & Kelley, H. H. (1953), 'Communication and persuasion.'.

Hovland, C. I. & Weiss, W. (1951), 'The influence of source credibility on communication effectiveness', *Public opinion quarterly* **15**(4), 635–650.

Iosif, E. & Mishra, T. (2014), From speaker identification to affective analysis: A multi-step system for analyzing children's stories, *in* 'Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)', pp. 40–49.

Knobloch-Westerwick, S. & Kleinman, S. B. (2012), 'Preelection selective exposure: Confirmation bias versus informational utility', *Communication research* **39**(2), 170–193.

Krippendorff, K. (2018), *Content analysis: An introduction to its methodology*, Sage publications.

Küçük, D. & Can, F. (2020), 'Stance detection: A survey', *ACM Computing Surveys (CSUR)* **53**(1), 1–37.

Kumar, S., Kumar, G. & Singh, S. R. (2022), Detecting incongruent news articles using multi-head attention dual summarization, *in* 'Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing', pp. 967–977.

Landis, J. R. & Koch, G. G. (1977), 'The measurement of observer agreement for categorical data', *biometrics* pp. 159–174.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692* .

Long, Y., Lu, Q., Xiang, R., Li, M. & Huang, C.-R. (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)', pp. 252–256.

Lukasik, M., Cohn, T. & Bontcheva, K. (2015), Classifying tweet level judgements of rumours in social media, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', pp. 2590–2595.

Lukasik, M., Srijith, P., Vu, D., Bontcheva, K., Zubiaga, A. & Cohn, T. (2016), Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', pp. 393–398.

Mann, W. C. & Thompson, S. A. (1988), 'Rhetorical structure theory: Toward a functional theory of text organization', *Text-interdisciplinary Journal for the Study of Discourse* **8**(3), 243–281.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D. (2014), The stanford corenlp natural language processing toolkit, *in* 'Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations', pp. 55–60.

Marcu, D. (1999), 'Instructions for manually annotating the discourse structures of texts', *Unpublished manuscript, USC/ISI* .

McFarland, S. G. (1981), 'Effects of question order on survey responses', *Public Opinion Quarterly* **45**(2), 208–215.

McGrath, M. & Frank, D. (2018), Propositions, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2018 edn, Metaphysics Research Lab, Stanford University.

McHugh, M. L. (2013), 'The chi-square test of independence', *Biochemia medica* **23**(2), 143–149.

Metzger, M. J., Flanagin, A. J. & Medders, R. B. (2010), 'Social and heuristic approaches to credibility evaluation online', *Journal of communication* **60**(3), 413–439.

Meyer, P. (1988), 'Defining and measuring credibility of newspapers: Developing an index', *Journalism quarterly* **65**(3), 567–574.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *In ICLR Workshop Papers.* .

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. (2016*a*), A dataset for detecting stance in tweets, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)', pp. 3945–3952.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. (2016*b*), Semeval-2016 task 6: Detecting stance in tweets, *in* 'Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)', pp. 31–41.

Molek-Kozakowska, K. (2013), 'Towards a pragma-linguistic framework for the study of sensationalism in news headlines', *Discourse & Communication* **7**(2), 173–197.

Morante, R., Van Son, C., Maks, I. & Vossen, P. (2020), Annotating perspectives on vaccination, *in* 'Proceedings of The 12th Language Resources and Evaluation Conference', pp. 4964–4973.

Nelson, M. (2019), Propositional attitude reports, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2019 edn, Metaphysics Research Lab, Stanford University.

Newell, C., Cowlishaw, T. & Man, D. (2018), Quote extraction and analysis for news, *in* 'Proceedings of the Workshop on Data Science, Journalism and Media, KDD', pp. 1–6.

Newell, E., Margolin, D. & Ruths, D. (2018), An attribution relations corpus for political news, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)'.

Norman, G. (2010), 'Likert scales, levels of measurement and the "laws" of statistics', *Advances in health sciences education* **15**(5), 625–632.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I. & Honnibal, M. (2012), A sequence labelling approach to quote attribution, *in* 'Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning', Association for Computational Linguistics, pp. 790–799.

Pareti, S. (2012), A database of attribution relations., *in* 'Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC12)', pp. 3213–3217.

Pareti, S. (2015), Attribution: a computational approach, PhD thesis, The University of Edinburgh.

Pareti, S. (2016), Parc 3.0: A corpus of attribution relations, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 3914–3920.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R. & Koprinska, I. (2013), Automatically detecting and attributing indirect quotations, *in* 'Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing', pp. 989–999.

Park, K., Kim, T., Yoon, S., Cha, M. & Jung, K. (2020), Baitwatcher: A lightweight web interface for the detection of incongruent news headlines, *in* 'Disinformation, Misinformation, and Fake News in Social Media', Springer, pp. 229–252.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018), Deep contextualized word representations, *in* 'Proceedings of 2018 conference of NAACL: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.

Pomerleau, D. & Rao, D. (2017), 'Fake News Challenge Stage 1 (FNC-1): Stance Detection', `http://www.fakenewschallenge.org/`.

Pouliquen, B., Steinberger, R. & Best, C. (2007), Automatic detection of quotations in multilingual news, *in* 'Proceedings of Recent Advances in Natural Language Processing', pp. 487–492.

Prasad, R., Dinesh, N., Lee, A., Joshi, A. & Webber, B. (2006), Annotating attribution in the penn discourse treebank, *in* 'Proceedings of the Workshop on Sentiment and Subjectivity in Text', Association for Computational Linguistics, pp. 31–38.

Qazvinian, V., Rosengren, E., Radev, D. & Mei, Q. (2011), Rumor has it: Identifying misinformation in microblogs, *in* 'Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing', pp. 1589–1599.

Rajapakse, T. (2017), 'Simple Transformers', `https://github.com/ThilinaRajapakse/simpletransformers`.

Rangel, F., Giachanou, A., Ghanem, B. H. H. & Rosso, P. (2020), Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, *in* 'CEUR Workshop Proceedings', Vol. 2696, Sun SITE Central Europe, pp. 1–18.

Rath, B., Salecha, A. & Srivastava, J. (2022), 'Fake news spreader detection using trust-based strategies in social networks with bot filtration', *Social Network Analysis and Mining* **12**(1), 1–19.

Reich, Z. (2011), 'Source credibility and journalism: Between visceral and discretional judgment', *Journalism Practice* **5**(1), 51–67.

Revilla, M. A., Saris, W. E. & Krosnick, J. A. (2014), 'Choosing the number of categories in agree–disagree scales', *Sociological Methods & Research* **43**(1), 73–97.

Riedel, B., Augenstein, I., Spithourakis, G. P. & Riedel, S. (2017), 'A simple but tough-to-beat baseline for the fake news challenge stance detection task', *arXiv preprint arXiv:1707.03264* .

Roberts, C. (2010), 'Correlations among variables in message and messenger credibility scales', *American behavioral scientist* **54**(1), 43–56.

Roy, A., Fafalios, P., Ekbal, A., Zhu, X. & Dietze, S. (2022), 'Exploiting stance hierarchies for cost-sensitive stance detection of web documents', *Journal of Intelligent Information Systems* **58**(1), 1–19.

Rubin, V. L., Chen, Y. & Conroy, N. J. (2015), Deception detection for news: three types of fakes, *in* 'Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community', American Society for Information Science, p. 83.

Ruths, D. (2019), 'The misinformation machine', *Science* **363**(6425), 348–348.

Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E. & Palomar, M. (2021), Exploring summarization to enhance headline stance detection, *in* 'International Conference on Applications of Natural Language to Information Systems', Springer, pp. 243–254.

Shin, K.-Y., Song, W., Kim, J. & Lee, J.-H. (2019), News credibility scroing: Suggestion of research methodology to determine the reliability of news distributed in sns, *in* '2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 737–740.

Silverman, C. (2015), 'Lies, damn lies and viral content', *Tow Center for Digital Journalism* .

Slovikovskaya, V. & Attardi, G. (2020), Transfer learning from transformers to fake news challenge stance detection (FNC-1) task, *in* 'Proceedings of the 12th Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 1211–1218.

Somasundaran, S. & Wiebe, J. (2010), Recognizing stances in ideological on-line debates, *in* 'Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text', Association for Computational Linguistics, pp. 116–124.

Soni, S., Mitra, T., Gilbert, E. & Eisenstein, J. (2014), Modeling factuality judgments in social media text, *in* 'Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', pp. 415–420.

Stab, C. & Gurevych, I. (2014), Identifying argumentative discourse structures in persuasive essays, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 46–56.

The House of Commons, U. (2018), 'Disinformation and 'fake news': Interim report'.

Wang, W. Y. (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', pp. 422–426.

Wardle, C. & Derakhshan, H. (2017), 'Information disorder: Toward an interdisciplinary framework for research and policymaking', *Council of Europe Strasbourg* .

Wathen, C. N. & Burkell, J. (2002), 'Believe it or not: Factors influencing credibility on the web', *Journal of the American society for information science and technology* **53**(2), 134–144.

Wertgen, A. G. & Richter, T. (2020), 'Source credibility modulates the validation of implausible information', *Memory & Cognition* **48**(8), 1359–1375.

Wiebe, J., Wilson, T. & Cardie, C. (2005), 'Annotating expressions of opinions and emotions in language', *Language resources and evaluation* **39**(2), 165–210.

Xi, N., Ma, D., Liou, M., Steinert-Threlkeld, Z. C., Anastasopoulos, J. & Joo, J. (2020), Understanding the political ideology of legislators from social media images, *in* 'Proceedings of the international aaai conference on web and social media', Vol. 14, pp. 726–737.

Yeung, C. Y. & Lee, J. (2017), Identifying speakers and listeners of quoted speech in literary works, *in* 'Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)', pp. 325–329.

Yuxi Pan, Doug Sibley, S. B. (2017), 'Fake News Challenge - Team SOLAT IN THE SWEN', `https://github.com/Cisco-Talos/fnc-1`.

Zhang, Q., Liang, S., Lipani, A., Ren, Z. & Yilmaz, E. (2019), From stances' imbalance to their hierarchicalrepresentation and detection, *in* 'The World Wide Web Conference', pp. 2323–2332.

Zhang, Q., Yilmaz, E. & Liang, S. (2018), Ranking-based method for news stance detection, *in* 'Companion Proceedings of the The Web Conference 2018', pp. 41–42.

Zhang, S., Zheng, D., Hu, X. & Yang, M. (2015), Bidirectional long short-term memory networks for relation classification, *in* 'Proceedings of the 29th Pacific Asia conference on language, information and computation', pp. 73–78.

Zibran, M. F. (2007), 'Chi-squared test of independence', *Department of Computer Science, University of Calgary, Alberta, Canada* pp. 1–7.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018), 'Detection and resolution of rumours in social media: A survey', *ACM Computing Surveys (CSUR)* **51**(2), 32.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R. & Lukasik, M. (2016), Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations, *in* 'Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers', pp. 2438–2448.

Appendices

Appendix A

Error Analysis: Detecting Stance Using Attribution Relation

A.1 News article: Negation in the headline

Headline:

Sorry, Argentina's President Didn't Actually Adopt a Jewish Werewolf

News body:

The President of Argentina, Cristina Fernandez de Kirchner, has adopted a Jewish godson – to prevent him from becoming a werewolf.

Although this sounds like something straight out of a fantasy novel, the President last week met Yair Tawil and his family for the unusual ceremony, which dates back more than 100 years and is based on Argentinian folklore.

According to the legend, the seventh son of a family will transform into 'El Lobison', a werewolf like creature, on the first Friday after the boy's 13th Birthday, and will continue to turn into a blood-thirsty, baby eating werewolf every full moon.

The fear of the creature was so fervent in 19th century Argentina that many families murdered or abandoned their seventh born son, forcing the Argentinian government to implement the process of Presidential adoption.

The tradition was established in 1907, and was extended to baby girls in 1973.

Although having seven children is now much rarer than 100 years ago, seventh sons or daughters can now expect to gain the President as their official godparent, as well as a gold medal and educational scholarship.

The President has said that Yair is the first Jewish boy to take part in the ceremony, as the tradition was exclusive to Catholic children until 2009.

A.2 News article: Fake cues in the headline

Headline:

Report: Woman who claimed to have the third breast added is fake

News body:

A Florida massage therapist said she paid $20,000 for a third breast in hopes of becoming less attractive to men

"I don’t want to date anymore," Jasmine Tridevil told Orlando’s Real Radio 104.1.

Tridevil, 21, has documented her post-surgery life through photos and videos posted to YouTube, Facebook and other social media sites – mostly images of her posing in custom-made three-cup bikinis and bras.

Her desire to repel the opposite sex with her updated anatomy wasn’t her only motivation for the surgery: Tridevil also hopes to have her own show on MTV someday.

She said she contacted more than 50 doctors before she found a surgeon willing to perform the operation.

A.3 News article: Preventive word in the headline

Headline:

Saudi Airlines to ban gender-mixing seating

News body:

According to an airline source quoted in stories claiming the opposite, Saudia Airlines, the state-run airline of Saudi Arabia, does not have plans to separate passengers based on gender.

A.4 News article: Opposing word in the headline

Headline:

Jasmine Tridevil: Woman with three breasts denies surgery hoax claims

News body:

SHE made headlines around the world when she revealed she paid thousands of dollars to get a third breast surgically attached to her chest.

But now Florida woman Jasmine Tridevil is facing claims that the surgery is a fake and she made it all up.

The claims come as an American news channel 10 News revealed Tridevil, who goes by the real name Alisha Jasmine Hessler, had filed an incident report after losing a three breast prosthesis earlier this month.

Tridevil filed the report after her bag was stolen from Tampa International Airport on September 16, with the report also obtained by TMZ.

The luggage, which was stolen, allegedly contained a fake breast and 10 News obtained the report which listed the missing contents including a "3 breasts prothesis".

However, Tridevil claims she underwent her surgery a few months ago and it cost $20,000.

Internet rumour site Snopes also uncovered her JasmineTridevil.com site is registered to Alisha Hessler, someone who Tridevil bares striking resemblance to.

However when 10 News reporter Charles Bill tracked her down, Tridevil insisted her surgery and third breast were the real deal.

"I figured people would be sceptical, but it's true. I recorded the surgery and it will be on my show," she said.

Tridevil insisted the surgery went ahead and she tried 50-60 doctors before she found one willing to perform the surgery.

Surgeons have also dismissed the possibility of it being real.

New York plastic surgeon Matthew Schulman told The Daily Dot: "[I] believe 100 per cent that this is a hoax that everyone is falling for," he said.

"I would be happy to go on record claiming that this is a falsified story and essentially not possible."

Michigan surgeon Dr Anthony Youn agreed while the surgery was possible, it was highly unlikely anyone would perform it.

In a YouTube clip Tridevil said the extra breast felt like her other two but that "the only difference is the nipple", which she had to get tattooed on.

The 21-year-old saved up for two years so she could have the surgery and is also paying for a film crew to follow her around.

She also said her surgery was documented by a film crew and will prove her story is true, The Mirror reported.

Tridevil has received plenty of venom on social media and just hours ago posted "Pain is temporary, glory is forever" from her Facebook page.

Just an hour earlier she wrote: "Don't be afraid to be different ... that's what makes you beautiful" and also posted a photo of her pre surgery which won her a host of compliments and questions as to why she would want to change.

"Omg ... I done know what is happening but I'm seriously going viral right now,"

It's not the first time Ms Hessler has made headlines.

Last year, she chose to publicly humiliate a man who beat her instead of sending him to prison.

She said she was introduced to the man when friends of hers invited him out clubbing last December and he beat her after unwanted sexual advances.

She received hospital treatment and a police report was filed.

But instead of pressing charges she offered her attacker an ultimatum telling him, "I can either press charges and have you arrested for a year, or I can have you sit outside at a busy intersection for 8 hours holding up a sign that says 'I beat women'."

A.5 News article: 1-AR news bodies

Headline:

Report: White House Chief Of Staff Denis McDonough: No Threats Were Made To Foley, Sotloff Families Over Possible Ransom

News body:

The US threatened to prosecute James Foley's family over ransom payments.

A.6 News article: Headlines with no claims

Headline:

Giant Crab

News body:

Nowadays, it seems the world is just littered with crises. But one thing we don't have to worry about is "Crabzilla."

A photo depicting what appears to be a 50-foot crab hanging out near a wharf in Kent, England, started making the rounds on the Internet earlier this week. The photo was featured on a local curiosities site called Weird Whitstable, and eventually dubbed "Crabzilla."

Quinton Winter, who runs the site, told the Daily Express that he'd spotted a giant crab in the mouth of Kent harbor while on vacation with his son last year. "It had glazed blank eyes on stalks, swiveling wildly and it clearly was a massive crab with crushing claws," he said.

"Does this satellite photo of the harbor reveal a giant crab or unusual sand formation?" Winter asks on his website.

Experts say neither.

"The idea of a giant 'crabzilla' would [be] very exciting. Unfortunately, I think this is a hoax," Dr. Verity Nye, Ocean and Earth Science researcher at Southampton University, told the Daily Mail. "I don't know what the currents are like around that harbor or what sort of shapes they might produce in the sand, but I think it's more conceivable that someone is playing about with the photo."

Spider crabs, the largest known to British waters, grow to about 4 feet and inhabit much deeper waters than the those near the pier where Crabzilla is shown lurking.

Appendix B

Annotation Guidelines

Background:

Attribution relation is a process of attributing an object such as a piece of text to its respective source/speaker. Formally, an attribution relation is a composition of three components which are as follows:

- source: A communicative agent
- cue: A lexical anchor that connects the content to the source. The cue expresses source's knowledge, attitude or intention towards someone or something.
- content: Part of text that is attributed to the source

Example: For the following text,

> Computational linguist Dr. Mathew Regg said to the team that
> the model has a good performance for text processing.

the attribution relation is as follows:

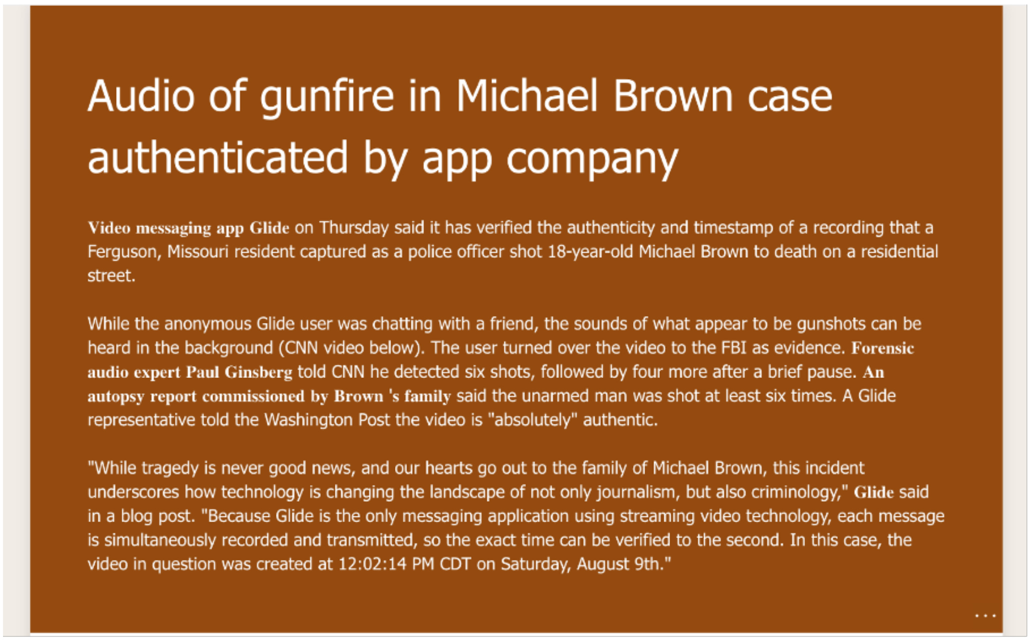
> source= Computational linguist Dr. Mathew Regg
> cue= said
> content= the model has a good performance for text processing

Annotation Instructions:

1. Please read the headline at the top of each annotation form and the

following article body of the news.

Following is a screenshot from an annotation form showing a headline and its respective article body.



Audio of gunfire in Michael Brown case authenticated by app company

Video messaging app Glide on Thursday said it has verified the authenticity and timestamp of a recording that a Ferguson, Missouri resident captured as a police officer shot 18-year-old Michael Brown to death on a residential street.

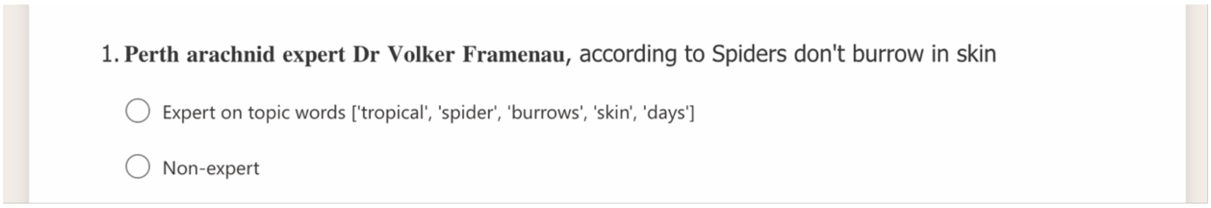
While the anonymous Glide user was chatting with a friend, the sounds of what appear to be gunshots can be heard in the background (CNN video below). The user turned over the video to the FBI as evidence. **Forensic audio expert Paul Ginsberg** told CNN he detected six shots, followed by four more after a brief pause. **An autopsy report commissioned by Brown 's family** said the unarmed man was shot at least six times. A Glide representative told the Washington Post the video is "absolutely" authentic.

"While tragedy is never good news, and our hearts go out to the family of Michael Brown, this incident underscores how technology is changing the landscape of not only journalism, but also criminology," **Glide** said in a blog post. "Because Glide is the only messaging application using streaming video technology, each message is simultaneously recorded and transmitted, so the exact time can be verified to the second. In this case, the video in question was created at 12:02:14 PM CDT on Saturday, August 9th."

Figure B.1: A screenshot showing a news headline-body pair

In the article body, the boldfaced texts are the sources.

2. After the news article in the annotation form, there are numbered article text snippets which are attribution relations extracted using a computational model. Following is a screenshot from an annotation form showing an attribution relation.



1. **Perth arachnid expert Dr Volker Framenau,** according to Spiders don't burrow in skin

○ Expert on topic words ['tropical', 'spider', 'burrows', 'skin', 'days']

○ Non-expert

Figure B.2: A screenshot with source boldfaced in an attribution relation

Here, the source is **Perth arachnid expert Dr. Volker Framenau**, the cue is *according to* and the content is *Spiders don't burrow in skin*.

In the annotation form, an attribution relation consists of a boldfaced source followed by a cue and then a content. The content is the claim made by the source.

3. Please annotate the source as an expert or non-expert considering the given topic words. Figure B.2 is a screenshot from an annotation form showing topic words suggestion.

   In the Figure B.2, considering the given topic words [‘tropical’, ‘spider’, ‘burrows’, ‘skin’, ‘days’], annotators should decide if the source Perth arachnid expert Dr. Volker Framenau is an expert or not. The purpose of providing the topic words is to help annotators to decide the domain expertise of the source.

   Please consider topic words with regard to the headline rather than treating each topic word as a domain. In the given screenshot, as the news headline is Tropical spider burrows under man’s skin, lives there for three days, a tropical fruit expert may not be considered as a domain expert in the given context.

   The topic words are derived from the headline using a computational model. The topic words in the annotation forms vary as per the news headlines.

   The following points can also be helpful to decide the source’s domain expertise:

   - The source with its domain name and “expert” explicitly mentioned in the article can also be considered as an expert source. For example: Forensic audio expert Paul Ginsberg

   - The source expertise can also be found according to their given domain names such as mineralogist, astronomer and so on.

   - The source expertise can also be decided from their duration of contribution in a domain, if mentioned in the article. For example: Dr. Matt, who spend 20 years researching on the insects

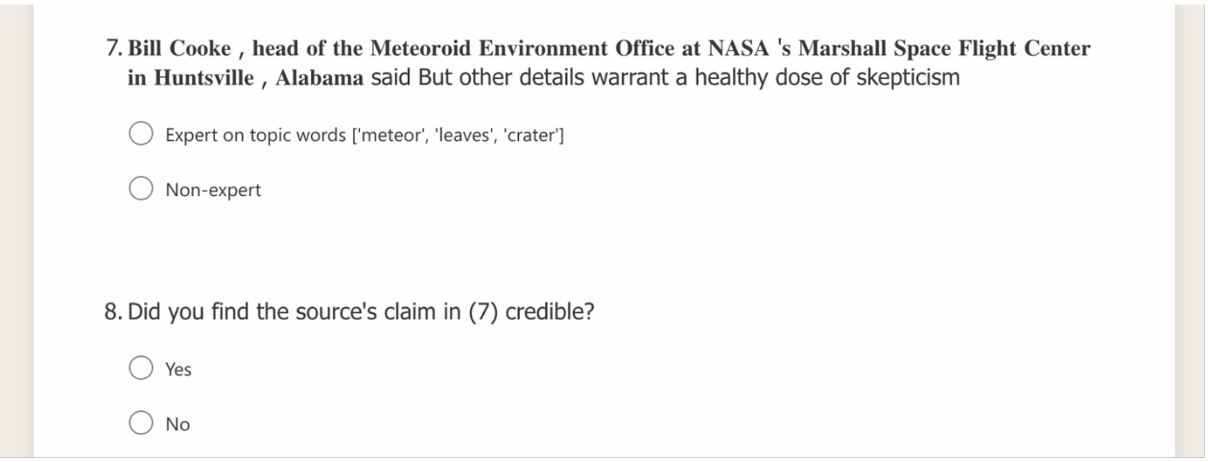
   - The association with a domain expert organization and the source’s position can also be other features for source expertise decision. For example: Hepburn of the U.K. Gout Society

4. A source annotated as an expert in a news article remains an expert throughout the article even though that source is represented by just

its name, a pronoun or any additional information.

Example:
The source **Computational linguist Dr. Mathew Regg** if considered as an expert of language processing, he is considered as the same in the news article no matter if he is mentioned as **Dr. Regg**, **Mathew** or **he**.

5. After the domain expert decision, a second question is further asked for each attribution relation.
Following is a screenshot showing the second question for an attribution relation.



7. **Bill Cooke , head of the Meteoroid Environment Office at NASA 's Marshall Space Flight Center in Huntsville , Alabama** said But other details warrant a healthy dose of skepticism

○ Expert on topic words ['meteor', 'leaves', 'crater']

○ Non-expert

8. Did you find the source's claim in (7) credible?

○ Yes

○ No

Figure B.3: A screenshot showing a further question asked for the claim credibility

As shown in the screenshot, Question (8) asks about the perceived message credibility of the claim *But other details warrant a healthy dose of skepticism* in Question (7). The perceived message credibility can be decided based on the source's expertise and the claim's relevancy to the headline. Please answer the second question for each attribution relation.
We asked the second question to analyse whether the source expertise is helpful to assess the credibility of the claim (or perceived message credibility).

6. Please submit the form after completion.

Thank you :)

Appendix C

Selected news articles for the survey

News Article 1:

Headline:

Expert casts doubt on Bunbury man Dylan Thomas's burrowing

News body:

Prepare to meet ... mite man. Doctors removed a matchhead-sized insect, believed to be a spider, from under Dylan Thomas's skin earlier this week and have sent the creature away for testing to determine what it is. It had been there for three days and burrowed up to his chest, leaving a trail of red blisters. The 21-year-old was on his first trip to Bali. He told News Corp yesterday that doctors had pulled a tropical spider "a bit bigger than the size of a match head" from his skin. There's just one problem. Spiders, according to Perth arachnid expert Dr Volker Framenau, don't burrow in skin. "They don't have the tools, the armature, to do this sort of stuff," Dr Framenau said. "I find it highly unlikely, almost impossible, that it was a spider.' More likely, Dr Framenau said, was some kind of burrowing mite. "That's a professional skin-digger," he said. "There's a lot of nasty stuff out there." The results of the tests on the creature should come back next week. Mr Thomas has been contacted for comment.

News Article 2:

Headline:

Meteorite Leaves House-Sized Crater in Nicaragua's Capital

News body:

A blast near the Nicaraguan capital city of Managua on Saturday night was most likely caused by a meteorite plummeting to Earth, creating a 40-foot-wide crater. A piece of the 2014 RC asteroid that passed close to Earth on Sunday, the meteorite dug a 16-foot-deep hole, Nicaraguan government scientists said. Miraculously, no one was hurt. The 60-foot-wide asteroid was passing 25,000 miles from Earth on Sunday but posed no danger to the planet, NASA said. Authorities have yet to determine whether the meteorite is buried or whether it disintegrated when it hit the ground. At first, locals believed the blast was caused by an earthquake, a regular occurrence in the country, Reuters said. "All the evidence that we've confirmed on-site corresponds exactly with a meteorite and not with any other type of event," said Jose Millan of the Nicaraguan Institute of Earth Studies. The explosion on the outskirts of Managua, near the city's airport, took place at around 11 p.m. and left a crater the size of a house. "It could have come off that asteroid because it is normal for that to occur. We have to study it more because it could be ice or rock," said Humberto Garcia, a Nicaraguan volcanologist. Calling it a "fascinating event," Nicaragua's first lady, Rosario Murillo, said the country would work with the U.S. Geological Service to find out more about what happened, The Daily Telegraph said.

News Article 3:

Headline:

Priest's claim of seeing God as a woman dismissed as hoax

News body:

A supposed Catholic priest's claims of seeing God as a woman when he died for 48 minutes are being described as a hoax. The Sun reported a man known as Father John Micheal O'Neal was declared dead by doctors at the Massachusetts General Hospital, near Boston, after suffering a massive heart attack but awoke 48 minutes later claiming he had seen God. "Her presence was both overwhelming and comforting," he said. "She had a soft and soothing voice and her presence was as reassuring as a mother's embrace. The fact that God is a Holy Mother instead of a Holy Father doesn't disturb me." But the Church has poured Holy Water on the claims. Terrence Donilon, a spokesman for the Archbishop of Boston, told Metro.co.uk they had no record of O'Neal being a priest. "We do not have a priest of this name. I believe this could be a hoax story."

News Article 4:

Headline:

Batmobile wasn't stolen: Cops

News body:

Rumours that the Caped Crusader's ride has been stolen have been greatly exaggerated. On Friday, bleedingcool.com said, "The scuttlebutt from sources in Detroit is that one of the Batmobile models being used in the filming of Batman Vs. Superman has gone missing, believed stolen." Not surprisingly, the Internet went into a tizzy, but later that day, Detroit police said the theft was a rumour. Sgt. Michael Woody told the Detroit Free Press that police confirmed with producers of Batman v. Superman: Dawn of Justice that the vehicle has not been stolen. "The Batmobile is safe in the Batcave where it belongs," Woody said. The paper also said that sources close to the movie being filmed in D-Town also said the fly ride had not been stolen. Unauthorized photos of the Batmobile appeared online this week, and director Zack Snyder tweeted an official photo on Wednesday. Batman v. Superman stars Ben Affleck and Henry Cavill, and is scheduled to open in theatres in 2016.

Appendix D

Most frequent cues in the training and test set of the FNC-1 corpus

D.1 30 most frequent attributional cues in the Training set of the FNC-1 corpus

cue	count
say	32473
tell	7734
accord to	6361
report	3718
claim	2618
show	1514
say :	1466
add	1429
have say	938
confirm	922
write	919
announce	736
be	689
be report	687
believe	682
describe	664
be say	659
say in a statement	596
know	560
also say	538
appear	503
note	481
reveal	439
tweet	429
be believe	417
declare	391
explain	374
call	373
ask	366
:	366

D.2 30 most frequent attributional cues in the Test set of the FNC-1 corpus

cue	count
say	10447
accord to	3930
tell	3607
report	2165
claim	1328
say :	670
show	577
add	528
ask	491
believe	472
write	468
appear	458
describe	387
also say	368
be	332
be say	325
be believe	320
reveal	314
explain	312
be report	283
suggest	282
confirm	276
seem	275
insist	267
have say	267
list	259
note	253
think	225
state	217
threaten	204

D.3 20 most frequent attributional cues per stance class in the Training set of the FNC-1 corpus

agree	**disagree**	**discuss**
say	say	say
tell	tell	tell
accord to	accord to	accord to
report	claim	report
claim	report	claim
say :	write	add
show	say :	show
write	add	say :
confirm	show	have say
add	note	confirm
say in a statement	think	be
believe	explain	write
announce	have report	be report
have say	be say	announce
ask	have say	describe
be say	be	be say
know	confirm	believe
describe	also say	also say
explain	announce	know
state	estimate	appear

D.4 20 most frequent attributional cues per stance class in the Test set of the FNC-1 corpus

agree	**disagree**	**discuss**
say	say	say
tell	accord to	accord to
accord to	tell	tell
report	report	report
claim	claim	claim
say :	show	say :
ask	write	appear
add	insist	also say
show	list	believe
explain	seem	show
write	explain	add
describe	describe	be believe
believe	say :	be say
think	reveal	write
be	ask	suggest
reveal	find	have say
insist	appear	be report
find	point out	confirm
state	want	ask
appear	read	describe