Corpus-based and Computational Analysis of Entity Framing

Esther van den Berg

Department of Computational Linguistics
Heidelberg University

This dissertation is submitted for the degree of
Doctor of Philosophy

Reviewer: Prof. Dr. Katja Markert
Second reviewer: Prof. Dr. Andreas Witt

Submission date: 01.11.2022.

Acknowledgements

For her patience and kindness, her tireless efforts, her principles, and her decision to choose me for a venture into Computational Social Science, I thank my supervisor Katja Markert. I will always be grateful that you continued to have faith in me through some very hard times, and that we made it through the strange pandemic years to see the completion of this thesis.

The work on Twitter data from this thesis would not have been possible without the painstaking efforts of the resourceful Katharina Korfhage, the annotations of my friend Juri Opitz, and the guidance and help of Josef Ruppenhofer and Michael Wiegand. For their corrections and feedback on this lengthy document, I thank my second supervisor Andreas Witt and my dear friends Mahsa Vafaie and Erinn Soest. I also thank Antonina Werthmann, Sandra Suchowitz and Cornelia Conrad for helping navigate the intricacies of three different contracts.

I am thankful to Saif Mohammad, Amber Boydstun and Claudia Wagner for inspiring this work and shaping my thoughts on the topic of entity framing with their talks and research. This thesis was written as part of the LiMo Leibniz-Wissenschafts campus, and benefited from interactions and meetings with Ines Rehbein, Vivi Nastase, Maria Becker, Bich Ngoc Do and Bhushan Kotnis. My colleagues Todor Mihaylov, Ana Marasovic, Debjit Paul and Juri Opitz helped make this work possibly with interesting talks, emotional support, great hikes and great dinners. The same can be said of Federico Lopez, Jeff Kuntz and Vera Nienaber. Special thanks go to Ipek Baris, Indira Sen and Kevin Wunsch for exciting opportunities to connect with other institutes.

Sincere thanks go to my parents Henk en Karin van den Berg and to Job van den Berg en Joy Uriot, for supporting me unconditionally from a distance. Finally, there is no part of this thesis that was not somehow improved by the support of my husband, Dr. José Angel Daza Arévalo.

Abstract

Entity Framing is the selection of aspects of an entity to promote a particular viewpoint towards that entity. Compared to issue framing, it has received little attention in Framing research, and it has also received very little attention in Natural Language Processing (NLP). We investigate Entity Framing of political figures on social media and the news through the selection of objectively verifiable attributes like name, title and background information.

Despite indications that they signal the perceived status of the target and/or perceived solidarity with a target, naming and titling have not previously been quantitatively examined in terms of their relation to stance. In this thesis, we collect English and German tweets mentioning prominent politicians and show that naming variation relates positively to stance in a way that is suggestive of a framing effect mediated by respect. We show on the German corpus that this positive relation is impacted by differences in political orientation.

Having provided the first quantitative evidence for the relation between stance and the mentioning of the objectively verifiable attributes name and title, we turn to engineering efforts towards automating detection of the selective mentioning of background information. By nature, whether certain information constitutes an instance of Entity Framing depends on the context in which that information is provided. Nevertheless, previous work on detection of framing through background information has not explored the role of context beyond the sentence. We experiment with computational methods for integrating three kinds of context: context from the same article, context from other publishers' articles on the same event, and inclusion of texts from the same domain (but potentially different events). We find that integrating event context improves classification performance over a strong baseline. We additionally show through a series of performance tests that the improvement over this baseline holds specifically for instances that are likely to be more difficult to classify, as one would expect from a performance boost that is due to leveraging context.

The result of these studies is a collection of new data sets, methods and findings with respect to Entity Framing, contributed in the hope that further computational research on this topic will be conducted, in order to improve our understanding of Entity Framing in general and of political figures in particular.

Table of Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Research Questions . . . 4

1.3 Contributions . . . 5

1.4 Thesis Overview . . . 7

1.5 Published Work . . . 9

2 Background **11**

2.1 Framing Theory . . . 11

2.2 Machine Learning Fundamentals . . . 21

3 Related Work **37**

3.1 Introduction . . . 37

3.2 Framing Detection . . . 38

3.3 Subjectivity Analysis and Detection . . . 45

4 Framing Through Naming and Titling **49**

4.1 Introduction . . . 49

4.2 Corpus Construction . . . 51

4.3 Correlation between Stance and Naming Formality . . . 61

4.4 Impact of Political Orientation and Target Gender . . . 66

4.5 Discussion . . . 72

4.6 Summary and Future Work . . . 73

5 Detection of Informational Entity Framing **75**

5.1 Introduction . . . 75

5.2 The BASIL Corpus . . . 77

5.3 Sentence-only Baseline . . . 78

5.4 IEF Detection with context beyond the sentence . . . 81

6 Performance Analysis of Context-Inclusive Entity Framing Detection 93
6.1 Introduction . . . 93
6.2 Sentence Length . . . 96
6.3 Subjective Language . . . 96
6.4 Target . . . 98
6.5 Polarity . . . 99
6.6 Direction . . . 100
6.7 Part of Quotation . . . 100
6.8 Political Leaning . . . 101
6.9 Conclusion . . . 102

7 Conclusions and Future Work 103

Appendix A List of News Accounts 107

Appendix B Redacted BASIL Example 111

Appendix C Training Parameters 113

Appendix D BASIL IEF Targets by Commonness 115

List of Figures 119

List of Tables 121

List of Terms and Abbreviations 125

References

Chapter 1

Introduction

1.1 Motivation

Each reference to a person in text is the result of a number of choices. How prominently they are featured, what reference form is used for them, what role they play in the context: all these are choices a writer makes when they make reference to another person. Communicators have limited time and space to convey their message and must select those wordings that tell the clearest and most engaging story. Their story leaves an impression on the reader. This impression can lead to a change in the reader's opinion of the people that play a role in that story.

We know from the study of framing that even small changes in the way writers tell their stories can have a large effect on readers' perception (Entman, 1993; Scheufele, 1999; An and Gower, 2009; Chong and Druckman, 2007; Matthes, 2009). Formally defined, framing is the process whereby the selection and presentation of aspects of a topic promote a particular interpretation of that topic (Entman, 1993). A convincing example of the effect of framing is a study in which respondents were asked their opinion about permitting a radical hate group to hold a political rally (Sniderman and Theriault, 2004). If the question was prefaced with the phrase "given the importance of free speech", respondents were overwhelmingly in favor of permitting the rally. If the question was instead prefaced with the phrase "given the risk of violence", a majority of respondents was willing to ban the rally from taking place. Even more surreptitiously, 65% of a sample of U.S. respondents believe more should be spent on "assistance to the poor", but only 20% are willing to increase spending on "welfare" (Rasinski, 1989).

Framing is slowly gaining attention in the Natural Language Processing (NLP) community, which studies how computers can process, analyse and label large amounts of text and speech. As of yet, the focus has been primarily on using computational methods to analyse and detect the framing of topics. Such topics include international conflicts, immigration and smoking (Berinsky and Kinder, 2006; Card et al., 2015; Tsur et al., 2015). With the exception of a handful of publications (Card et al., 2016; Rashkin et al., 2015; Fan et al., 2019), little attention has gone to studying the effect of framing on opinions towards entities.

NLP tools have been developed to capture explicit bias and partisan language, and these may suffice to capture sentiment-bearing connotations of references to entities. An aspect of framing that they will miss, however, is that framing can occur through the selective mentioning of purely objectively verifiable attributes of entities. Consider, for example, the following case. According to pro-Israeli lobbyists, the BBC displayed bias in their reporting when they called all hostages in the Entebbe plane hijacking in 1976 "Israeli"[1]. While many of the hostages were indeed citizens of the state of Israel, some of them were Jewish citizens of other countries. Omitting this detail, detractors claimed, would impact perception of the hijackers' motivations. A hostage-taking of Israelis could be interpreted as a purely political, anti-zionist attack, whereas a hostage-taking of Jews reveals anti-semitic motivations. It is relatively straight-forward to imagine how these and other omissions (or inclusions) of attributes such as an entities' nationality, name, age, gender or profession could change the interpretation of events. What is not straight-forward, is how an NLP system that detects sentiment would flag such omissions or inclusions as relevant to sentiment.

The aim of this thesis is to venture into this relatively new terrain of the study of framing through objectively verifiable attributes. We will refer to this terrain as Informational Entity Framing (IEF), and define it as 'the selection or omission of objectively verifiable aspects of an entity to promote a particular viewpoint towards that entity'. Examples of IEF include inclusion or omission of demographic information like age, religion and nationality, biographical information like current and previous professions, and mentions of previous actions taken by the entity. They may also include quotations of statements the entity has made, or even quotations of comments and opinions of third parties about the target entity.

Our contributions in this area are two-fold. Firstly, we complement existing evidence of the impact of the framing on views towards topics with support for impact of IEF on opinions towards entities. Specifically, we present evidence that stance is significantly positively related to the formality of naming forms in novel labeled data sets in English and German. Secondly, we experiment with computational methods of detecting IEF and show the benefit of modeling context in the form of other news articles when detecting IEF in news. As part of this computational experiment, we conduct a series of tests that suggest that the performance increase from modeling context is not a generic improvement in training processes, but a specific improvement in classifying instances that are difficult to classify out of context.

[1] https://honestreporting.com/bbc-erases-non-israeli-jews-entebbe-hostages/

Because of the special role that framing in news media plays in influencing the public's world view and political attitude (Baumgartner et al., 2008; De Vreese, 2004; DellaVigna and Gentzkow, 2010; McCombs and Reynolds, 2009, 2002), we focus both contributions on text with political themes and consider only political figures as entities. We look at two text types: online discourse in the form of social media messages, and traditional newspaper texts. Online political discourse has been said to have an increasing influence on the democratic process (Ott, 2017; Perloff, 2021). Social media messages on political themes have also been shown to be retweeted more often when they contain negative appraisals of political parties and figures (Dang-Xuan et al., 2013). Political news reports, meanwhile, require journalists to describe newsworthy events as neutrally and objectively as possible. Because they nevertheless have to contextualise these events and structure their reporting in such a way that readers can interpret the information meaningfully, implicit subjectivity in the form of framing may be particularly common in this text type, and detection particularly important.

Because units of text on social media platforms and news media have very different average lengths, we do not study the same types of objectively verifiable attributes in each domain. In our experiment with social media texts, we study an objectively verifiable attribute that can be expressed in very few characters, and that can be recognised in incomplete or ungrammatical utterances: the name and title of the political figure in question. Names and titles of entities convey information about the perceived status of the entity and about perceived solidarity with the entity (Brown and Gilman, 1960). We are interested in demonstrating that variation in this attribute contributes to the subjective meaning readers derive from texts that mention political figures. Because no labeled data sets exist for this purpose, we will construct novel data sets labeled for naming form and stance.

In our experiment with news texts, we build on top of work which has shown that news texts frequently frame entities by including tangential, speculative or background information that reflects negatively or positively on these entities (Fan et al., 2019). This may include descriptions of their past, quotations of opinions from their allies or opponents, or allusions to possible reasons for their actions. Since background information can take the form of a phrase or an entire sentence, the detection of IEF through background information is a free-form task and therefore exceptionally challenging. We are motivated to experiment with computational approaches to this task despite its challenging nature because of the increasing influence of NLP on journalism. Journalists are beginning to use NLP tools like fact-checking and quote verification software (Ali and Hassoun, 2019; Morris et al., 2021), and the amount of automatically produced news is also on the rise (Túñez-López et al., 2021). We anticipate that efforts are needed from within the research community to ensure that AI tools for monitoring the neutrality of human and machine-generated texts are designed to flag not only potentially incorrect facts and biased wording, but also neutrally formulated, factually correct text with framing potential. To summarize, the motivation of this thesis is to provide evidence of the impact of entity framing through names and titles on social media on the perception of political figures, and to improve the detection of informational framing of political figures through background information in news texts.

1.2 Research Questions

The previous section provided our reasons for studying IEF through naming and titling on social media and through background information in news texts. Here we break this goal down into three distinct Research Questions.

1.2.1 Analysing Entity Framing Through Naming and Titling

As stated, our work is motivated by the belief that entity framing in public discourse impacts public attitudes towards political figures. While there is some evidence that framing influences attitudes towards policies and topics, it has not been quantitatively established whether this holds for attitudes towards entities as well. We aim to gather evidence in support of the notion that variation in the use of names and titles contributes to the subjective meaning of a social media message. We investigate whether names and titles are likely to subtly encourage a certain opinion or perception of an entity in social media messages mentioning political figures. To establish a connection between names and titles on the one hand and opinion on the other, we ask whether a quantitatively measurable relationship exists between naming and titling and the perceived stance of a text in a sufficiently large and varied sample of social media data in which political figures are mentioned by name. Our first research question is thus:

Is there a measurable relation between naming/titling and stance in social media messages that mention politicians?

The effect of naming and titling on perceived stance may vary in nature and effect depending on sociolinguistic context, because different communities may assign different connotative meanings to names and titles. It has been suggested that progressive and conservative groups differ in how they value respecting and maintaining status differences (Haidt and Joseph, 2004), which suggests they may assign different connotations to the use of names and titles to signal such differences. We suspect that any relation between naming and stance discovered to answer the previous research question will differ in direction and/or strength depending on whether the data used to measure this relation is produced by progressive or conservative social media users. There is also some evidence that titles are more likely to be used when addressing males than females (Takiff et al., 2001). We therefore also expect to find observable difference in the use of names and titles for male as opposed to female politicians. We test both suspicions by answering our second research question, which is:

Do naming and titling patterns in social media messages differ depending on the language of the messages, the political leaning of the source and the gender of the target of a reference to a political entity?

1.2.2 Improving Detection of Informational Entity Framing by Including Context

Our first and second research question are designed to gather information about the impact of a minor change in choice of words, to show that the inclusion of objectively verifiable attributes can impact readers' interpretation of texts' stance towards entities. However, in the envisioned target use case of framing detection systems for news writers, the challenge is not only to recognise individual lexical choices, but to recognise spans of non-essential background information of varying lengths. Little work has been done to attempt to solve this problem, and the work that has been done has focused on recognising framing through background information in isolated sentences. By definition, background information can only be assessed as having a framing effect in relation to a foregrounded element. Whether a piece of information is essential to introducing an entity or its role in an event depends on the content of the other pieces of text describing the same event, be it other sentences in the same article or even text from other articles. We thus propose to improve detection of IEF by modeling potential candidates of IEF together with their context. Our research question with respect to improving detection performance for IEF is thus as follows:

Does inclusion of context from the same or other articles improve detection of IEF in news?

1.3 Contributions

The research questions described in Section 1.2 consist of two analytical questions and one engineering question. Because of this marked difference between the first two and the final research question, this dissertation is made up of two distinct parts which present research conducted with the distinct methods. The contributions of the dissertation, therefore, also come in two categories.

Our corpus-based analysis of framing through naming and titling in online political discourse makes the following contribution towards the study of entity framing:

1. We conduct the first quantitative study on the relation between the use of names and titles and the perceived stance of a text.

2. We provide the first two data sets annotated for both naming variation and stance, in two languages and for two sets of naming forms: the English Twitter Titling Corpus of roughly 4000 English-language tweets mentioning six presidents, and the German Twitter Titling Corpus of almost 2000 German tweets mentioning 24 members of the German parliament.

3. We provide evidence that names and titles, while not sentiment-bearing themselves, contribute to the perceived stance of a text.

4. We quantify previously made claims by sociolinguistics concerning the status-indicating function of naming that were thus far only based on small-scale and/or qualitative observations.

5. We provide evidence that the way in which names and titles contribute to stance differs between sociopolitical groups in a way that corresponds to differences in their value systems.

6. We show that, contrary to what one might expect, social media messages do not necessarily display gender bias in terms of the frequency with which titles are applied to female vs male politicians.

Our study of the impact of context-inclusion on IEF detection performance makes the following contributions towards improving IEF detection:

1. We conduct the first study of the impact of including different types of context on IEF detection.

2. We compare IEF detection systems in a novel setting that prevents information leakage from test into train data.

3. We are the first to attempt sequential sentence classification and extended pre-training of foundational models to IEF detection.
4. We present a simple method for data curation in preparation of extended pre-training.
5. We show that, contrary to what findings from previous work suggest, extended pre-training of foundational models on data sampled from the same distribution of the task is ineffective for IEF detection in news.
6. We present a novel context-inclusive method that models article and event context, the latter of which outperforms a strong sentence-only baseline and achieves state-of-the-art results in IEF detection for English on the BASIL corpus (Fan et al., 2019).
7. We demonstrate that the differences in performance between our proposed event context-inclusive model and the baseline are not evenly distributed across types of instances, but follow a pattern of improved performance for difficult cases across six dimensions of difficulty.

1.4 Thesis Overview

The chapter following this introduction, **Chapter 2**, provides the background required for understanding both the related work sections and the methods used in our computational study. It begins with a history and overview of Framing Theory research. This history covers the origins of the term *framing* in psychology, its subsequent adoption in other fields like Communication Science, Artificial Intelligence and Linguistics research, and its first appearances in NLP. We then disambiguate the term *framing* from other related terms and offer a taxonomy of types of framing to contextualise IEF. The final section on Framing Theory reviews the evidence for the impact of framing on readers' attitudes and decisions. The second half of Chapter 2 introduces fundamental concepts for computational modeling. We cover unsupervised machine learning methods that are common in related work, as well as supervised machine learning techniques that are relevant for understanding the systems we apply to IEF detection.

In **Chapter 3** we first briefly discuss the abundance of terms for various kinds of implicit subjectivity in NLP literature, and establish which terms will be used with what meaning from thereon. We then give a review of relevant prior work on the detection of framing, where we cover both detection of the framing of topics and of entities. The chapter continues with related research in a related but distinct line of work which focuses on the classification of bias. We also describe the similar but distinct work on the detection of explicit expressions of opinion and attitude. Finally, we cover studies by sociolinguists which investigate the function of forms of address which underpin the hypotheses regarding framing through naming and titling that are investigated in the subsequent chapter.

Chapter 4 presents our corpus-based analysis of framing of political figures through naming and titling. We motivate and describe the collection and annotation of our English and German data sets of social media posts that mention global leaders and German parliamentarians, respectively. We then present our central hypotheses regarding the likely relations between stance and naming and titling, which are informed by the sociolinguistic research presented in Chapter 3. Statistical analysis of our data reveals a relation between naming formality and stance that confirms previous sociolinguistic claims across two different time frames, languages and titles. We go on to analyse how the function of naming that we discovered to be dominant in our data sets differs in strength depending on the political orientation of the social media user, and the gender of the target politician.

Chapter 5 opens with evidence of the challenge that IEF currently poses to the NLP community by performing sentence-level IEF on a more sound data split that shows lower performance than previously reported results. This sentence-level baseline experiment is followed by a series of experiments with various types of context beyond the sentence. The types of context that are included are: context from neighbouring sentences, context from the same article, context from other articles on the same event, and context from other articles of the same domain. We correspondingly design and test the performance of techniques for encoding sequences of sentences, for encoding sentences from the same article as well as articles from other events, and for encoding the properties of the genre of news texts. Our findings demonstrate the importance of event context in particular for determining whether a sentence is an instance of entity framing.

Chapter 6 more closely analyses the performance of the best-performing context-inclusive IEF detection system to assess whether its performance gain reflects a benefit of access to context during training, or an overall improvement related to differences in the training process. The chapter takes six properties of candidate sentences that are likely to make them more difficult to classify and compares the performance of the best baseline and the best-performing context-inclusive IEF system on sentences that exhibit these properties to various degrees. The dimensions of difficulty that we consider include length of candidate sentences, presence of subjective language in candidate sentences, the target of the IEF-containing sentences, the polarity of IEF-containing sentences, whether IEF in sentences refers to targets directly or indirectly, whether IEF-containing sentences are part of a quote, and finally the leaning of the article of candidate sentences. For each dimension, we assess whether the best-performing context-inclusive IEF system outperforms the baseline by a larger margin on the more difficult subset of the data when data is grouped according to that dimension. We demonstrate that our best-performing context-inclusive IEF system outperforms the baseline by a larger margin for difficult data subset for six out of seven properties. This provides evidence in support of the notion that exposure to context helped the proposed system make better predictions in cases where the candidate sentence is difficult to classify out of context.

In **Chapter 7**, we draw conclusions from the presented findings and provide an outlook of possible directions for future work.

Appendix B gives an example from the BASIL (Fan et al., 2019) corpus used for the experiments in Chapters 5 and 6. **Appendix C** gives details about the training procedure for the models in Chapter 5.

1.5 Published Work

A large portion of the research presented in this thesis is an extension of published works of which the thesis writer was the first author. The English Twitter Titling Corpus (ETTC)[2] was presented in the Third Workshop on Natural Language Processing and Computational Social Science (van den Berg et al., 2019). This publication reports on the collection and annotation of tweets which mention six G20 presidents for the purpose of a corpus-based study of the potential framing effect of the use of first names, last name and the title *President*. The study compares tweets that use the title *President* with tweets that do not, in terms of their perceived stance towards the target entity. A positive correlation was found between naming formality and stance, which confirms sociolinguistic claims about the function of names and titles and indicating the status of the target.

The German Twitter Titling Corpus (GTTC)[3] was presented along with findings on the relation between naming/titling and stance in a separate publication as part of the Proceedings of The 12th Language Resources and Evaluation Conference (van den Berg et al., 2020). In this corpus of tweets mentioning German parliamentarians with a doctoral degree, we found a positive correlation between more formal naming, including use of the title *Dr.*, and stance. This confirms the aforementioned sociolinguistic claims about the status-indicating function of naming in a language other than English and for a title other than President. Additionally, we demonstrated that an interaction exists between naming formality, stance and the likely political orientation of the Twitter user who posted the tweet. We found that the positive correlation between naming formality and stance is weaker in tweets from left-leaning

[2] https://doi.org/10.11588/data/IOHXDF

[3] https://doi.org/10.11588/data/AOSUY6

users than right-leaning users. This finding constitutes empirical evidence from the area of computational sociolinguistics in support of the hypothesis from Moral Foundation Theory (Graham et al., 2009) that left-leaning and right-leaning users assign different importance to maintaining social hierarchies.

Our exploration of four kinds of context-inclusion in IEF detection was published as part of the Proceedings of the 28th International Conference on Computational Linguistics (van den Berg and Markert, 2020). In this paper we explored and proposed methods that include neighbouring sentences (referred to in this paper as direct textual context), sentences from the same article (article context), sentences from articles on the same event (event context) and sentences from other news articles on other events (domain context). We applied Sequential Sentence classification to model direct textual context, a Context-Inclusive Model to model article and event context, and extended pre-training of a language model to model domain context. The best-performing model was the Context-Inclusive Model with event context. This paper also contains a brief error analysis of which Chapter 6 is an extension. This analysis showed that including context benefits performance more on longer sentences, and sentences from articles with a centrist political stance.

Chapter 2

Background

2.1 Framing Theory

The analysis and detection of framing in NLP is a young research area that originated from a much larger one: Framing Theory (FT). FT's history is decades long and knows influences from many disciplines. In this sub-chapter, we provide a brief history of how framing research developed since the initial use of the term framing, and how the different categories of researchers that conduct framing research differ in their understanding of the concept.

Because this dissertation focuses on the framing of politicians on social media and in the news, its introduction to FT prioritizes the concept of framing as it is understood in Communication Science. In Communication Science, framing is considered to be an important factor in media influence on public discourse, closely related to other media effects like agenda-setting and priming. In fact, there exists debate within the field whether and how framing should be separated from these other terms. Because there is debate over the definition of framing, we take time towards the end of this section to consider ways of distinguishing between framing and related terms. We close with a description of the progress that has been made towards proving and characterising the influence of framing on public opinion.

2.1.1 Brief History of Framing Theory

Academic use of the term *framing* can be traced back to the psychiatrist Gregory Bateson (Bateson, 1955). Bateson wished to understand how humans convey to each other whether their actions are intended as play or not. He suggested that humans use meta-communication to frame actions, that is, to apply an interpretative framework to them, in this case the frame of play. Bateson understood framing as the invocation of temporary and spatial boundaries within which gestures take on a meaning different from the meaning they have outside of those boundaries. The term frame was used as an analogy to physical picture frames (Bateson, 1973). Like picture frames, social frames establish boundaries, focus attention on that which lies within them, and exclude aspects of reality that lie outside of those boundaries. Framing can drastically change the interpretation of an object. A cropped photograph may give a very different impression of an object than the same object seen within its full context. Within the frame of play, acts of aggression that would normally express hostility bring about relaxation and bonding.

In his milestone work *Frame Analysis*, sociologist Ervinn Goffman suggested that *all* human behaviour, not just play, can be understood in terms of frames (Goffman, 1974). Individuals in a society learn to recognise a catalogue of frames that specify what actions and reactions can be expected in a given situation. These frames are ubiquitous and simmer below our conscious awareness. Frames in this understanding are the psychological and sociological means by which humans organise and structure reality in a useful and socially coordinated way.

Goffman's concept of frames as a tool for structuring reality was adopted in psychology, Artificial Intelligence (AI) and linguistics. In AI, the term *frame* denotes a mental unit of knowledge akin to a script (Schank and Abelson, 1977) that humans and, by extension, artificially intelligent agents rely upon to categorize and navigate daily life (Minsky, 1974). A classic example of such a unit is the sequence of steps that are needed to order and pay for food at a restaurant. In linguistics, frames form the cornerstone of the theory of semantics developed by Charles Fillmore (Fillmore et al., 1976). In this theory, frames are scenarios invoked by verbs that have been internalised as scripts to more efficiently navigate recurring experiences. Returning to the example of paying for food, the conceptual structure or frame behind placing an order might consist of a verb like *pay* and the semantic roles of the person paying, the item that is being paid for, and the recipient of the payment (Fillmore et al., 2006).

Where psychologists and semanticists focused on framing as an internal process by which humans structure experiences, sociologists began to investigate the role of frames in communication. They used the term *framing* to describe the way journalists and other communicators portray events to their audiences (Tuchman, 1980; Gitlin, 2003). The notion of framing as the selective foregrounding or omitting of aspects of news events found fertile ground in Communication Science, where it met the closely related theory of first-level and second-level agenda setting. Agenda setting theory posits that the media influences the public by selecting which topics and which aspects of topics are to receive the most attention (McCombs and Shaw, 1972, 1993). To distinguish framing from other terms and unify differing interpretations of the term, Entman (1993) offered a definition that is still cited often. It states that framing is to “select some aspects of a perceived reality and make them more salient in a communication text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation” (Entman, 1993, p. 52).

Entman’s definition of framing mentions selectivity: framing foregrounds aspects of reality instead of other aspects that are equally real. Later work emphasises that the media does not select aspects at random, but in line with a communities’ values (Van Gorp, 2007; Scheufele and Tewksbury, 2007). The selected aspects are chosen, often subconsciously, to fit common ways of structuring events. These common ways of structuring events may be general or culturally specific. For example, to help readers engage with and remember a news event, a journalist might frame it as a hero overcoming adversity, and depending on the audience, they may add a reference or metaphor that evokes scenes from superhero comics.

The first calls to automatically detect media framing according to Entman’s definition can be found in Greene and Resnik (2009), Choi et al. (2012) and Recasens et al. (2013). Researchers in the field of NLP have made various attempts to automatically analyze framing in text that continue to attract new studies (Card et al., 2015; Tsur et al., 2015; Fan et al., 2019; Akyürek et al., 2020). Recent work has even contributed novel terms to Framing Theory, including the concept of *subframes* (Roy and Goldwasser, 2020) (subtypes of a generic frame, e.g. a Minimum Wage Economy frame and Salary Stagnation frame as expressions of the Economy frame) and of *informational bias* (framing of entities through tangential, speculative or background information) (Fan et al., 2019). Section 3.2 discusses the study of framing within NLP where it relates to IEF in more detail.

2.1.2 Delineating Framing

Throughout the history of the study of framing, the terms *framing* and *frames* have been applied to a great variety of phenomena. Framing has been used to denote the change of semantic framework through meta-communication (Bateson, 1955), the organisation of lived experiences into schemas that guide behaviour (Goffman, 1974; Minsky, 1974), to the selective describing of aspects of reality by journalists (Tuchman, 1980; Entman, 1993; Scheufele and Tewksbury, 2007; Tankard et al., 1991). Even excluding the psychological and semantic concepts of frames and focusing on the concept of framing in the media, there is a great variety of phenomena that can and have been called framing.

The heterogeneity of approaches to framing has raised concerns among scholars of media framing. Communication Science research seeks to understand effects on public opinion caused by media. The lack of a unified theory and methodology of framing is seen as an obstacle that impedes systematic progress towards that goal (Entman, 1993; Scheufele and Iyengar, 2012). Concerns exist that existing definitions of framing are not concrete enough to facilitate empirical research, and that those empirical methods that have been attempted are not accepted broadly enough to enable comparisons between findings (Weaver, 2007; Chong and Druckman, 2007).

In the remainder of this section, we cover common sources of confusion. There is, as stated, no universally accepted definition and taxonomy of framing, but this section describes commonly recognised subcategories of framing and reflects consensus in related work on how framing is distinct from similar concepts where it exists. The intention here is not to provide a comprehensive overview of each of the concepts discussed, but to instead give clarity about which understanding of framing this thesis builds upon.

2.1.2.1 Framing vs. Schemas

In Communication Science, and in this dissertation, when we refer to frames we are referring to frames in communication. We can call frames in communication **media frames** to set them apart from frames in thought, which we can also call **schemas**. Media frames are repeated combinations of small components that are used in public communication (Van Gorp, 2007). They are informed by the current culture and shaped by the needs and constraints of the medium, such as the need to keep attention over the duration of a news report or to constrain a message to a two-dimensional image or text (Scheufele and Tewksbury, 2007; Kinder et al., 1996; Scheufele, 1999; Brewer, 2003). Schemas, on the other hand, are constructs which guide the interpretation of messages once they reach the receiver. They exist only in the individual's psyche and may be unique to an individual (Scheufele and Tewksbury, 2007).

2.1.2.2 Framing vs. Agenda Setting

An important source of concern surrounding the definition of framing stems from its strong resemblance to **second-level agenda setting**. Agenda setting theory posits that the media influence public discourse by selecting, firstly, which topics and events to attend to (first-level agenda setting), and secondly, which aspects of those topics and events should be given the most attention (second-level agenda setting) (McCombs and Shaw, 1972, 1993).

To some, framing and second-level agenda setting are equivalent concepts. If a distinction must be made, *framing* refers to a special case of agenda-setting where micro-attributes band together to form a frequently invoked macro-attribute that resonates particularly well with audiences (McCombs and Ghanem, 2001).

According to others, agenda setting and framing should be seen as complementary rather than synonymous. In this view, agenda setting focuses primarily on the selection of aspect of topics, while framing focuses on the presentation of those selected aspects (Price et al., 1997; Scheufele and Tewksbury, 2007). Agenda-setting research studies the *what* of topic aspect selection. It examines the distribution of topics and aspects of topics in the news along with their perceived relevance to news consumers. The goal is learn how attention and repetition in the media influences the salience of topics or their attributes in the eyes of the public (Scheufele, 2000). Taking a new environmental policy aimed at reducing carbon emissions as an example, an agenda-setting study might assess which aspects of the policy are covered with what frequency, and notice a tendency to highlight economic implications in a subset of outlets. It might then investigate whether, over time, survey respondents with exposure to these outlets rated concerns regarding economic growth as a reason to question the policy more frequently than other respondents.

Framing research, meanwhile, focuses on the *how* of topic aspect selection. It goes beyond measuring the degree of attention to delve into the nature of that attention. It may also examine how these presentations form larger narratives, as well as how they shape attitudes, beliefs and even voting behaviour. Returning to the example of a new environmental policy, a framing approach might investigate whether possible negative economic impacts were presented by sceptical outlets in the form of statistics from financial institutes, quotes from experts, conversations with concerned business owners, or other means. It may then go on to examine how these presentation strategies relate to the values, fears and priorities that characterise these outlets and the audiences they target.

2.1.2.3 Framing vs. Priming

Also closely related and sometimes confused with framing is **priming**. This Communication Science concept refers to the observable impact of framing in the media on the standards by which the public judges public officers (Iyengar et al., 1987; Miller and Krosnick, 1996)[1]. For example, frequent attention to freedom of speech will prime audiences to base their overall approval of a politician in terms of that figure's perceived competence with respect to protecting freedom of speech. In the context of this thesis, we consider a change or consolidation in standards for evaluation a consequence of framing, and we consider research on priming equivalent to research on the impact of framing on public opinion and reception of political figures and their policies.

[1] This meaning of the term is distinct from the meaning of the term in psychology, where it refers to a stimulus that precedes a target stimulus and increases the speed with which the target stimulus is retrieved from memory. Research suggests these two types of priming do not refer to the same process (Miller and Krosnick, 1996)

2.1.2.4 Framing vs. Persuasion and Propaganda

Where framing refers to selective emphasis in the presentation of an issue, **persuasion** refers to a process of changing opinions. Framing is but one of several possible strategies to achieve persuasion (Hiebert, 2003; Da San Martino et al., 2019). Conversely, persuasion need not be the intended outcome of framing. Framing may be done intentionally to change a certain attitude or behaviour in an audience, but it could also intend to reinforce existing beliefs. Alternatively, it can simply be a recurring way of approaching a topic in media with no assumed intention on the part of the communicators.

A similar distinction can be made between framing and **propaganda** (Moloney, 2006). Unlike framing, propaganda is associated with a clear intention, usually on the part of governing powers, to instill or solidify a viewpoint in the audience, with framing being a possible strategy for reaching this goal.

2.1.3 Types of Framing

2.1.3.1 Equivalency Framing vs. Emphasis Framing

Emphasis framing, also known as issue framing, is framing as we have discussed it until now and as we will discuss it for the remainder of this dissertation, i.e. as selectively presenting aspects of an issue as more important and downplaying or omitting others. **Equivalency framing**, on the other hand, refers to a specific subcategory of framing through similarly structured, logically equivalent formulations of the same fact or proposition (Tversky and Kahneman, 1981). If emphasis framing is to discuss a policy in terms of its effect on employment or its effect on the environment, than equivalency framing is to present either a) a prospective 90% decrease in unemployment or b) a 10% increase in employment as result of the policy (Rasinski, 1989). Research has shown that such logically equivalent phrasings can strongly impact responses in survey settings, as we will discuss in more detail in Section 2.1.4 (Tversky and Kahneman, 1981; Sniderman and Theriault, 2004).

2.1.3.2 Generic Framing vs. Issue-specific Framing

It is a source of debate whether framing research should concentrate on finding abstract frames that can be applied across topics, time periods and cultures, or whether it should instead use coding schemes for finding frames that are tailored to the topic at hand (Entman et al., 2004; Chong and Druckman, 2007). Generic framing research is less common than issue-specific framing research. Examples of comparative studies based on generic frames include a study on the framing of crises of various types (An and Gower, 2009) and work on the Samsung Galaxy Note 7 explosion on tweets from three countries (Kang et al., 2019). Much attention has gone to the use of the generic conflict frame for coverage of politics, where war-related and game-like terms are used to describe political events like elections (Rhee, 1997; Zoizner, 2018). Although purely generic framing research is rare, many coding schemes use generic frames along with issue-specific frames. These are often the five generic media frames adapted by Semetko and Valkenburg (2000) from Neuman et al. (1992): 'attribution of responsibility', 'conflict', 'human interest', 'morality' and 'economic consequences'.

More commonly, framing studies start with the selection of an issue or event. Inductive issue framing studies describe patterns from direct observation of sampled reports on the topic. Deductive studies use a pre-defined coding scheme informed either by previous work on the selected topic, observations of communication by the elite, or frames elicited from individuals in interviews or surveys (Chong and Druckman, 2007). In deductive studies, categories can be mapped to a vocabulary of strongly associated terms (e.g. 'unemployment' for an 'economy' frame) which can then be marked in text automatically using content analysis software. Communication science studies that go beyond automatic string matching and use machine learning methods are discussed in Section 3.2. Recent issue framing research has focused on racial attitudes and representations of marginalised groups (Kellstedt, 2003; Noakes and Wilkins, 2002; Igartua et al., 2005), perceptions of international conflicts (Dimitrova et al., 2005; Berinsky and Kinder, 2006; Edy and Meirick, 2007) and the framing of scientific findings (Nisbet et al., 2003; Dahl, 2015). Popular material also includes crises of various types, including the Euro crisis (Kaiser and Kleinen-von Königslöw, 2019), public health issues (Lawrence, 2004), climate change (Kause et al., 2019) and the COVID-19 pandemic (Wicke and Bolognesi, 2020).

This thesis sits somewhere between a generic and an issue-specific approach. We do not take an issue-specific view, because we look into the framing of entities irrespective of the event or topic that is covered in the text, and we do not take an entity-specific view, because we aggregate findings and performance statistics across multiple entities in all of our studies. Instead, we take an entity-type-specific view and gather findings on the framing of entities that hold political offices.

2.1.3.3 Issue Framing vs. Entity Framing

Although most framing research is focused on **issue framing**, the framing of topics, events or issues, framing also applies to entities, in which case it may be referred to as **entity framing**, which is the focus of this thesis. Whenever entities are referenced in communication, the reference will convey both referential meaning, i.e. information about the characteristics of that entity, and indexical meaning: information about the entity's position in society, and their position with respect to the writer and the reader (Richardson, 2006; Blommaert et al., 2005).

One way in which these meanings are conveyed is through lexical choices. An example is the choice of describing a person who participates in an armed separatist movement as either a terrorist or a freedom fighter. Another example is supporters and opponents of an endangered species listing referring to the animals as either orcas or killer whales (Harden, 2006). Each verb chosen to describe an event also conveys not only information about what happened, but also selects a limited number of involved entities, specifies their role in the event, and often carries evaluative connotations about those roles (Rashkin et al., 2015). With the nouns and verbs they choose, writers thus intentionally or unintentionally assign roles to entities and position them with respect to each other.

Referential and indexical meaning may also be linguistically encoded. These linguistically encoded meanings have not been a topic of study in Framing Theory, but Section 3.2.2.3 describes observations from sociolinguists that point to possible framing effects of the semantics of terms of address.

Frequently, the assigned roles are drawn from a cast of stereotypes and cultural archetypes with strong connotations, like villain, tragic hero, or damsel in distress (Van Gorp, 2010; Card et al., 2016; Schneider and Ingram, 1993). Van Dijk (1998) observes that journalists may position entities in an **ideological square** to create a contrast between entities. In such a square, one positively represented actor is presented as an insider, and another negatively portrayed actor as an outsider. Role-assigning or stereotyping of entities can extend to entire groups, in which case there is undeniable potential for harm (Schneider and Ingram, 1993). That these patterns are likely to influence attitudes towards persons can be inferred from the ample evidence that news framing effects policy support and attitudes towards topics detailed in the following section (Baumgartner et al., 2008; De Vreese, 2004; DellaVigna and Gentzkow, 2010; McCombs and Reynolds, 2009, 2002).

2.1.4 Empirical Evidence of Framing Effects

When the first analyses of framing as a device in media communications appeared in academic writing on FT, it was uncertain whether framing had measurable effects on public discourse or behaviour. In this section, we give a brief overview of the progress that has been made since then towards demonstrating that through framing, communicators make connections between the topic at hand and existing opinions and values in the receivers mind and in doing so, measurably influence attitudes.

Most studies on framing effects ask whether frames used by politicians and in the media affect citizen's beliefs, opinions and behaviour. This question is hard to answer, in part due to a lack of widely accepted measures (Chong and Druckman, 2007), but some general observations have been made.

In one study, researchers found that framing appears to have a stronger effect when repeated in fairly rapid succession, rather than with longer intervals in between exposures to a frame (Lecheler and de Vreese, 2013). The strongest influence tends to be exercised by the most recently used frame, especially on participants with low levels of political knowledge (Lecheler and de Vreese, 2013). The presence and strength of a framing effect in this studies was determined by designing synthetic articles that containing descriptions of either economic opportunities or risks of entry of Bulgaria and Romania into the EU. Respondents were then exposed to either one of these articles or a neutral control stimulus. A framing effect was said to be present if respondents that had read the articles containing the economic frame reported opinions that were more aligned with those articles than respondents who read the control article.

Other evidence suggests that the proliferation and influence of a particular frame depends on the backing of that framing choice by lobbying camps, and that one can view media framing effects in terms of a competition between camps of advocates who fight with each other for dominance of their frame (Junk and Rasmussen, 2019). The degree to which frames resonate with common cognitive biases partially determines their effectiveness (Aarøe and Petersen, 2020). The group membership of the sender and receiver of the framed message also impact reception: frames appear to be less effective or even counter-productive when the communicator represents a political party that the receiver dislikes (Slothuus and De Vreese, 2010).

In addition to general observations regarding the influence of framing, there is a wide variety of studies on the influence of issue framing on opinions towards specific topics. A subset of these studies examine how framing sometimes restructures opinions, that is, how it changes the components of the reasoning behind an opinion without changing its polarity. Clawson et al. (2008) describe how white citizens expressed their support of a 1995 anti-affirmative action decision of the U.S. Supreme Court differently depending on how the decision was framed to them. If it was framed in terms of keeping government decisions separated from racial considerations, white supporters expressed that they were motivated by individualism. If the court decision was instead framed as a failure to address injustice, white supporters expressed motivation by resentment. In another study, framing of social welfare as either an undeserved handout or a drain on the economy did not change the degree of opposition to social welfare, but the former caused participants to focus on and express particular beliefs about poverty (Nelson et al., 1997).

Other evidence demonstrates actual changes of stance as a response to framing. Equivalency frames change the polarity of an opinion from favourable to opposing by changing the question that is used to elicit the opinion (Tversky and Kahneman, 1981; Sniderman and Theriault, 2004; Rasinski, 1989). In an example already provided in Section 1.1, respondents were polled for their approval of permitting a specific hate group to hold a political rally. If the question was prefaced with the phrase "given the risk of violence", only 45% of the respondents stated that they were in favor of allowing such a rally. If the question was instead prefaced with the phrase “given the importance of free speech”, the number of respondents in favor was 85% (Sniderman and Theriault, 2004). In another example, when asked whether more money should be spend on welfare, agreement jumped from 20% to 65% percent if “welfare” was called “assistance to the poor” (Rasinski, 1989).

Experimental studies on emphasis and issue framing outside of opinion polls have shown that here, too, framing can influence political opinions. For example, U.S. public support of military involvement in foreign conflicts was shown to be higher among readers of a news story describing it as a response to a humanitarian crisis, and lower for readers of an article which instead focused on risk of harm to military personnel (Berinsky and Kinder, 2006). Framing has also been shown to affect opinions towards government spending (Jacoby, 2000), influence evaluations of foreign nations (Brewer et al., 2003), increase levels of political cynicism (De Vreese, 2004) and contribute towards a long-term shift in attitudes towards the death penalty (Baumgartner et al., 2008)

In addition to influences on public opinion, framing also impacts public discourse in other ways. Politicians pay attention to framing when structuring their messages to the public (Jacoby, 2000; Zaller et al., 1992), and they react to the frames that are used in the media (Jacoby, 2000; Riker et al., 1996; Edwards and Wood, 1999; Zaller et al., 1992; Druckman et al., 2004). In turn, the media can be observed using issue frames that were first employed by politicians, or by other actors in public discourse such as lobbyists and activists (Scheufele, 1999; Entman et al., 2004; Carragee and Roefs, 2004; Fridkin and Kenney, 2005). There is also work which shows that citizens exchange and introduce frames to one another in every day conversations about political issues (Druckman and Nelson, 2003).

Finally, framing can have important effects outside of the realm of political discourse. Bibas (2004) showed that defendants make different decisions depending on whether they view plea bargains through a Loss frame or a Gain frame. When defendants mentally compare a plea bargain to acquittal, the bargain is seen through a Loss frame and experienced as a defeat, even though many defendants would benefit from accepting it. When defendants are kept in pre-trial detention for some time prior to deciding to a plea bargain offer, they can compare the plea bargain to imprisonment, and begin to view it more through a Gain frame, increasing the likelihood that they accept the offer. In another study, framing was found to influence patients' preferences for cancer treatments (O'Connor, 1989). Groups were presented with a toxic and a non-toxic treatment option, of which the former was more effective. Whether patients were willing to undergo the more effective, but toxic treatment depended on whether the option was framed to them in terms of the probability of surviving or in terms of the probability of dying.

Across these studies demonstrating the effects of framing, framing is preferably analysed by means of manual annotations with marginal assistance from software (Chong and Druckman, 2007). Manual analysis is more precise and thorough than computational approaches. However, it takes considerably more time and requires expert annotators. As evidence for the impact of framing mounts, it becomes desirable to speed up the annotation and analysis of framing using computational methods. Section 3.2 describes some attempts to do so. To understand these studies, the following section will explain the fundamental concepts which underpin current attempts to automate framing detection and analysis.

2.2 Machine Learning Fundamentals

Machine Learning (ML) is a branch of computer science that develops algorithms which can learn from data in ways that seek to imitate human pattern recognition capabilities. ML algorithms learn to produce accurate predictions for unseen data by tuning parameters of a mathematical model on a set of training data until an optimum is reached and the algorithm converges. In NLP, ML algorithms are used to solve language processing tasks. This includes labeling documents for the topics they feature, translating sentences from a source language to a target language, and determining the sentiment of mentions of brands on social media.

All ML experiments require the extraction of **features**: machine-readable representations of the training samples. In an NLP experiment, input features may be lists of numbers corresponding to indices in a lexicon that represent the words contained in an input sentence. Where ML methods differ is whether the model is trained using only these input features, or whether it also expects human-provided correct predictions. The former is called unsupervised learning, and the latter is called supervised learning.

In **unsupervised ML**, model parameters are fit to input features. An example of a common unsupervised algorithm is clustering, which treats inputs as points in a mathematical space of varying distance from one another, and compares the distances between data points to group them. Much framing research is completely manual, but a subset of framing studies use unsupervised algorithms to detect patterns in news texts that may shed light on framing strategies. In order to understand this partially automated framing research, we explain the foundations of unsupervised learning in Section 2.2.1.

In **supervised ML**, models are trained using input features along with corresponding human-provided examples of correct outputs, also referred to as **gold data**. In a Sentiment Analysis task, the gold data may be the ratings accompanying reviews from Rotten Tomatoes[2]. Supervised algorithms learn a mapping from inputs to outputs and optimize this mapping using a **loss function** which penalizes the system for deviating from the gold standard. How well the tuned model generalises to non-training data can be tested after training on held-out test data by computing values for performance metrics that assess the closeness of the model's predictions to the gold labels of the test set. The methods used in Chapter 5 of this dissertation build on supervised ML algorithms that are specialised in processing language and which are explained in Section 2.2.2.

2.2.1 Unsupervised Methods

Recognising subtle processes of communication like framing in texts is non-trivial and requires time and expertise to do successfully (Chong and Druckman, 2007). For this reason, framing researchers, like other social scientists, are increasingly turning to computational methods to replace or complement manual analysis (van Atteveldt and Peng, 2018). The method of choice is often an unsupervised algorithm, as this avoids the need for time-consuming manual annotation. Aside from being less costly than either manual content analysis or supervised learning, unsupervised methods also offer the advantage that they can easily be applied across topics, and scale well to large amounts of input data.

A closely related concept to unsupervised learning is self-training. Self-training minimizes the dependency on annotated data by producing pre-trained language models that learn lexical co-occurrence probabilities from raw texts. Self-training typically serves as a preliminary step to transfer learning, where a task-agnostic language model is fine-tuned using task-specific annotated data of a much smaller size than that required in traditional supervised learning. Because of its strong connection with transfer learning, self-training is explained in more detail in the section dedicated to transfer learning (Section 2.2.2.3).

2.2.1.1 Clustering

As the name suggests, clustering algorithms group data points together, based on measures of similarity between instances. The most well-known clustering technique is ***k*-means**

[2] `https://www.rottentomatoes.com/`

clustering (Lloyd, 1982). This algorithm is a hard clustering approach, which means that data points are assigned to one and only one cluster. It assigns data points to one of *k* possible clusters, where *k* is a number that is chosen a priori by the researcher. To group data points, one first transforms all of them into n-dimensional vectors, where each dimension i represents a feature of the data, and select k random points in this space as cluster centers. Next, each data point is assigned to the closest cluster center. The most commonly used distance metric is Euclidean distance: $D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where x_i and y_i are the values for feature i of data points x and y, and $\sum_{i=1}^{n}$ indicates the summation over all n features. Once the closest cluster center for each data point has been found, the third step is to set clusters' centers to be the mean of the vectors of all data points that got assigned to it. Step two and three are then repeated until the cluster centers no longer change substantially in between steps. At this point, the algorithm has converged and finishes. The k-means clustering process is illustrated in Figure 2.1.

Another common clustering technique is **Hierarchical Clustering** (Johnson, 1967). This is a bottom-up procedure that successively merges data points of a certain similarity together, while storing each merging decision, until all data points belong to one large cluster. To find the two closest clusters and merge them, so-called linkage formula are used that compute the distance between clusters in terms of the distances between data points within those clusters. For example, for single link clustering, a linkage formula is used that calculates the minimum of all distances between pairs of data points in two clusters: $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$ where C_i and C_j are two clusters and $d(x, y)$ is the distance between individual data points x and y. The two clusters with the lowest distance are merged, and this process is repeated until all data points belong to a single cluster. The user uses a heuristic or manual inspection to determine which intermediate step of the clustering procedure they wish to be made final. The advantage of this approach over k-mean clustering is that the final number of groups that is selected is informed by inspection of several possible options rather than being fixed before running the algorithm.

2.2.1.2 Topic Modeling

The assumption made by hard clustering that a data point can only belong to one category is a strong one that is often inappropriate for the task. A news report, for example, may belong to both the political and national news sections of a news site. In such cases, it is more appropriate to use a soft clustering approach. An example of a soft clustering approach that is used frequently in related work on Framing is **probabilistic topic modeling**. This algorithm returns not a group label, but the likelihood of a data point belonging to each of the possible categories. In probabilistic topic modeling, topic categories are treated as a latent

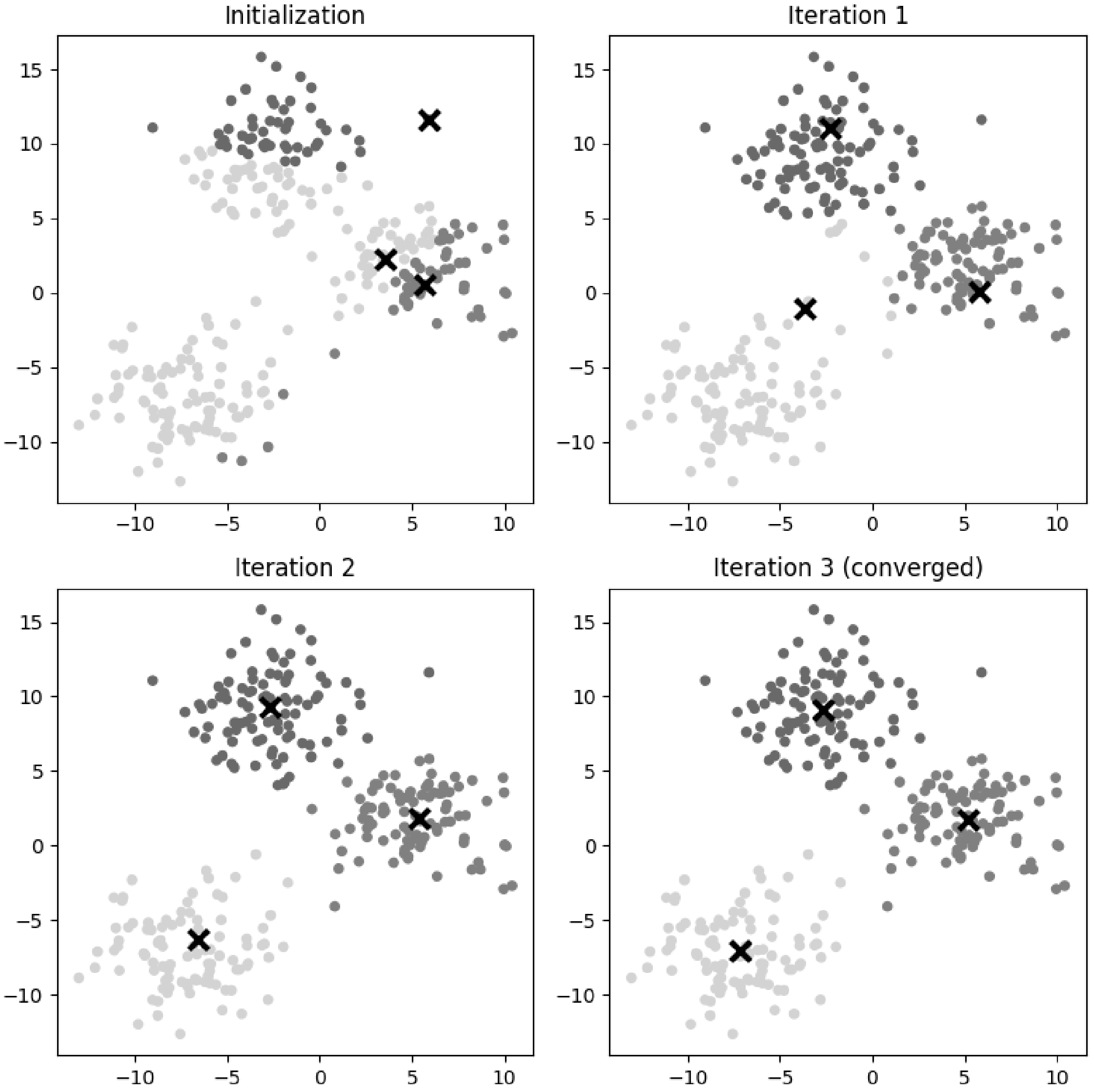

Fig. 2.1 Illustration of *k*-means clustering. Initalization: data points are projected in a vector space and cluster centers are initialized to random points in the space. Iterations 1-3: data points are iteratively assigned to the center closest to them and centers are then re-computed to be the mean of their data points until the algorithm converges.

variable. This means they are measured not directly from the data, but inferred from other measurements. More specifically, they are inferred as hidden themes from word-document co-occurrences in the input data. The result of this inference process is a generative model which specifies a probabilistic procedure by which documents can be generated by repeatedly drawing words with a likelihood determined by the distribution over topics. Bayesian statistical techniques are applied to this generative model to invert the generation process and compute which set of topics is most likely given an existing collection of documents.

The most widely used probabilistic topic models are based on **Latent Dirichlet Allocation** (Blei et al., 2003; Blei, 2012). LDA is the de facto standard model technique for topic semantics in text-as-data studies (Maier et al., 2018; Grimmer and Stewart, 2013). The algorithm assumes that each document d (newspaper article, tweet, or other unit of text) in a corpus D contains a mixture of topics from a predetermined number of topics K, where each topic is characterized by a distribution $beta_k$ over the vocabulary V containing the unique words in D (Blei et al., 2003, p. 996). LDA uses a Bayesian method of inference to determine the most likely topic structure $beta_k$, as well as the topic distributions for each document $theta_d$ from the observed word frequencies in D (Steyvers and Griffiths, 2007; Blei, 2012).

To perform LDA we start, as in *k*-means clustering, with the selection of a value for *k* specifying the desired number of possible topics K. The contents of the data are projected to a low-dimensional space and topic distributions for each topic $k \in \{1, \ldots, K\}$ are initalized from the Dirichlet distribution from which the approach derives its name: $\theta_d \sim$ Dirichlet(α). Topic assignments for each word in each document $n \in \{1, \ldots, N_d\}$ $z_{d,n}$ are initialized: $z_{d,n} \sim$ Multinomial(θ_d). Finally, a word w_d, n is drawn: $w_{d,n} \sim$ Multinomial$(\beta_{z_{d,n}})$. This generative step is necessary to update parameters.

During the inference phase, statistical techniques are used to update the topic assignments $z_{d,n}$ and topic distributions θ_d, as well as infer the word distributions $beta_k$ based on the observed documents. The inference phase is repeated until the algorithm converges. The goal is to maximize the log likelihood of the observed words given the topics and other parameters:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log P(w_{d,n} | z_{d,n}, \beta)$$

. Here $P(w_{d,n}|z_{d,n}, \beta)$ is the conditional probability of word $w_{d,n}$ given topic $z_{d,n}$ and word distribution β. A common implementation of LDA uses the Expectation-Maximization (EM) algorith to repeat the inference phase is repeated until convergence. A common follow-up step is to assign documents cluster labels corresponding to the topic with the highest value for $theta_d$.

LDA co-exists with a number of related techniques which includes **Structural Topic Models** and **Hierarchical Topic Models**. Structural Topic Modeling is an extension that enables the user to incorporate meta-data about documents. Examples include the name of the author or publisher of a text, or the date of publishing. This method has proven to be effective for analysing open-ended survey questions (Roberts et al., 2014). Hierarchical Topic Models have the advantage of providing an interpretation of the relationships between topics, and have been used to model a hierarchy of levels of framing (Nguyen et al., 2015) (see Section 3.2 for more detail).

2.2.1.3 Evaluation

Although unsupervised methods can be helpful to analyze large amounts of unlabeled data, they come with a set of challenges and limitations. Firstly, choosing an appropriate value for the *k* number of clusters or topics is highly impactful, but can be non-trivial to do reliably (Maier et al., 2018). Secondly, the outcome of unsupervised algorithms can be hard to interpret manually. Clusters and identified topics do not necessary correspond with the common-sense interpretation of the word *topics*, and there is no consensus in topic modeling literature on a theoretic definition for what the identified groupings represent (Günther and Domahidi, 2017). How most studies approach the problem of naming clusters is to identify the top words of each cluster, i.e. words with the highest probabilities assigned to them, and choose a label that describes the commonality between these words (Maier et al., 2018).

In addition to being hard to interpret, unsupervised learning results are also hard to compare to one another. If some amount of ground truth labels exist, there are metrics which can be used to assess the correctness of clusters (see e.g. Hubert and Arabie (1985)'s Rand index). If the output contains probabilities, a metric like **perplexity** can be computed on held-out data and used to assess how surprised the model is by data from a similar distribution to the training data. If no gold labels exist, quantitative evaluation may be performed by assessing which method scores better on desirable properties of the output such as topic coherence (the frequency with which top words co-occur in a cluster's documents (Mimno et al., 2011)), mutual information (the degree to which top words contribute significant information to a given topic (Grimmer, 2010)), and distinctness of clusters (how much closer points within a cluster are to each other than they are to points in nearby clusters (Rousseeuw, 1987)). Another option is to use external validation to increase confidence in the outputs by for example consulting experts (see e.g. Levy and Franklin (2014)) or using real-world events as anchors (see .e.g Evans (2014)). For more thorough discussions of the evaluation of unsupervised methods see the overviews of Maier et al. (2018) and Palacio-Niño and Berzal (2019)

2.2.2 Supervised Methods

Although supervised methods require labeled data, they are very effective and lend themselves better to comparative, performance-oriented experiments. Supervised methods can be divided into non-neural methods and neural methods. Traditional non-neural methods are necessary background literature for understanding the supervised work on framing that has been done in the text-as-data domain. Neural methods are crucial to understanding current NLP work and are likely to play a key role in future work on automating framing detection.

2.2.2.1 Non-neural Methods

Non-neural methods appear in text-as-data studies conducted by social scientists, and may also be used as a baseline to compare neural methods to.

One common and effective non-neural method is the **Naive Bayes (NB)** classifier (Bishop and Nasrabadi, 2006). Bayes' theorem defines the probability of an event based on prior knowledge of conditions that are assumed to impact that event as follows:

$$P(event|conditions) = \frac{P(event) * P(conditions|event)}{P(conditions)}$$

Where $P(event|conditions)$ is called the posterior probability, $P(event)$ is called the class prior probability, $P(conditions|event)$ is called the likelihood, and $P(conditions)$ is called the Predictor Prior Probability. A NB classifier uses as many copies of this rule as there are classes in a task to generate the probability that an instance belongs to each of these classes given the known values for features are assumed to impact class membership. The training algorithm solves the problem of learning the probability´s for each class from the input data, by assuming that features are independent from each other. This means that, for example, if trained to predict membership to groups with or without cardio-vascular disease from patients' height and weight, the model will learn a naive association between weight and the likelihood of having cardio-vascular disease that does not take into account that patients' weight is high or low relative to their height. Once all associations have been optimised, the model can be used as a classifier which assigns new instances to their most likely class. Despite its naive assumption of feature independence, and despite its simplicity, NB is in many cases a solid baseline for the more complex approaches that will be discussed in the next section.

Another common and powerful non-neural model is the **Support Vector Machine (SVM)** (Cortes and Vapnik, 1995). SVMs learn to automatically categorize data by calibrating a hyperplane in the feature vector space that efficiently divides data points into two groups. Efficient division here is defined as finding a hyperplane that is the furthest away from the data points on either side of the plane. The data points that are used as reference point for the plane are called its support vectors. Note that this division is binary: there is a single hyperplane dividing two groups. To divide the data into more than two groups, the algorithm can be repeated on the initial two sections, and again on subsequent divisions, until the desired number of groups is reached.

2.2.2.2 Neural Networks

Artificial Neural Networks (ANNs) or simply **Neural Networks (NNs)** have became immensely important in NLP due to their ability to model large numbers of dependencies and solve non-linear tasks, including complex tasks like machine translation, speech recognition and language modeling. NNs require large amounts of training data, but if this requirement is met, they require less feature engineering work and process more information than non-neural methods, which typically use sparse vectors that capture only some of the content of the input data. With the advent of transfer learning, which transfers basic knowledge extracted from huge unlabeled datasets to task-specific training experiments on smaller, labeled datasets, neural methods have become ubiquitous in any work involving the computational processing of language.

NNs derive their name from the biological neuron. Neurons in the human brain receive external input and transmit an electric signal to surrounding neurons if that external input triggers an activation. This conditional activation sets neurons apart from linear processes that always return output values that are proportional to their inputs. The insight that learning from examples could be made possible by neuron-like structures dates back to the 50ies, with the development of **perceptrons** (Rosenblatt, 1958; Minsky and Papert, 1969). Perceptrons have the following mathematical definition:

$$\mathbf{y} = f(\mathbf{xW} + \mathbf{b}) \tag{2.1}$$

where $\mathbf{x} \in \mathbb{R}^{d_{in}}$ is an input vector with dimensionality d_{in}, $W \in \mathbb{R}^{d_{in} \times d_{out}}$ is a matrix of weights; $b \in \mathbb{R}^{d_{out}}$ is a bias term and f is an activation function that maps input to output values $\mathbf{y} \in \mathbb{R}^{d_{out}}$. The activation function determines whether the neuron should be active or inactive given the input, for instance based on whether the input is greater or equal to zero.

When neurons with conditional activation are interconnected, they can learn complex patterns and perform non-linear classification. A stack of perceptrons is called a **Multi-Layer Perceptron** (MLP). In an MLP, information is passed between layers of neurons. A 3-layer MLP, for example, can be defined as:

$$NN(\mathbf{x}) = f''(f'(f(\mathbf{xW}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2)\mathbf{W}_3 + \mathbf{b}_3) \tag{2.2}$$

where inputs are transformed by matrix W_1 and bias b_1 through the activation function f. The transformed input then undergoes another transformation at the hands of matrix W_2 and bias b_2 and non-linear activation function f'. In the third and final layer, the twice transformed input is turned into output by matrix W_3, bias b_3 and f''. This passing along of transformed data lends simple MLPs of this type a second name: **Feed-Forward Neural Networks (FFNNs)**. To adapt FFNNs for classification, outputs are transformed linearly with a final matrix to k dimensions where k is the number of classes, and class membership is assigned to the class with the highest predicted value.

FFNNs were not adopted widely until the discovery of a mechanism for adjusting the values for weights and biases from training data. This mechanism, called *backpropagation*, involves a scoring of the distance between initial predictions and gold labels, followed by the application of *gradient descent* to adjust weights and biases in a direction that will improve this score (Rumelhart et al., 1985; Werbos, 1990). The invention of backpropagation accelerated research into neural architectures, which evolved into larger and more complex structures such as those described in the next paragraphs. To reflect the increase in the size and complexity of neural network structures, NNs began to be referred to as "deep", and neural machine learning came to be known as **Deep Learning**.

A drawback of FFNNs of any depth is their inability to retain information from previous instances. Each instance in the training data is processed on its own and related only to the gold label provided with it, not to previous instances or labels. This is a problem for the processing of language, as words in utterances do not exist in isolation, but have interdependencies with other words in the same utterance.

As a solution to this problem, networks called **Recurrent Neural Networks (RNNs)** were created that possess a mechanism to retain information over time (Elman, 1990). An RNNs is a function that takes as input an ordered sequence of vectors with dimension d_{in}, formally $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T], \mathbf{x}_t \in \mathbb{R}^{d_{in}}$, and returns a corresponding series of vectors with dimension d_{out}, formally $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T], \mathbf{o}_t \epsilon \mathbb{R}^{d_{out}}$. Every $\mathbf{o}_t$ represents not $\mathbf{x}_t$ in isolation, but summarizes the sequence up to $\mathbf{x}_t$ through a recursion mechanism. This entails that at each time-step *t* the RNNs updates a current hidden state $\mathbf{h}_t$ based on the inputs $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$, i.e. the instance and state previous to the current state, by applying the function:

$$\mathbf{h}_t = f(\mathbf{x}_t\mathbf{W} + \mathbf{h}_{t-1}\mathbf{U} + \mathbf{b}) \tag{2.3}$$

which consists of the feed-forward function Eq. 2.2.2.2 and a matrix $\mathbf{U}$ which acts as the network's memory by learning how best to use the previous hidden state $\mathbf{h}_{t-1}$. Once the entire input sequence has been processed, the output vectors $\mathbf{O}$ can be used as the input for a word-level labeling layer which predicts e.g. part-of-speech tags. Alternatively, the final output vector $\mathbf{o}_T$, which is identical to the final hidden state $\mathbf{h}_{t=T}$, can be considered an encoded representation of the input sequence. If the sequence $\mathbf{X}$ is a sentence (i.e. a sequence

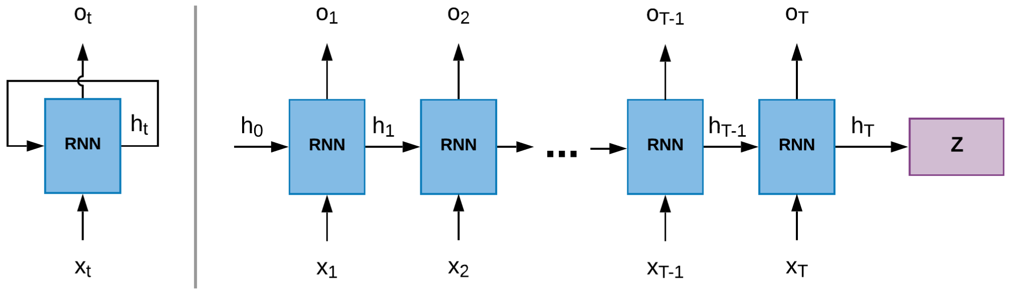

Fig. 2.2 Diagram illustrating a Recurrent Neural Network in two ways. On the left, the RNN is depicted in its folded state. On the right, it is unfolded through time to reveal as many states as items in the input sequence X. At each time step *t* the RNN takes as input a fixed-size vector $\mathbf{x}_t$ and a hidden state $\mathbf{h}_t - 1$, and outputs a vector $\mathbf{o}_t$ which represents $\mathbf{x}_t$ and a hidden state $\mathbf{h}_t$ which represents the sequence from $\mathbf{x}_1$ until $\mathbf{x}_t$. The final hidden state, $h_{t=T}$ is returned as $\mathbf{o}_T$ to represent $\mathbf{x}_T$ or as z to represent the entire sequence $\mathbf{x}_T$, depending on the task the system is used for.

of word-tokens), the final hidden state $\mathbf{h}_{t=T}$ can thus be treated as a representation of the whole sentence, denoted then as vector $\mathbf{z}$, and can be used for sentence-level classification. Figure 2.2 illustrates this property of the RNN architecture.

RNNs' weakness is that while they technically retain information of instances from the past, they in practice often lose information that is more than a few steps away. This is because of a problem referred to as the vanishing gradient problem. Small values multiplied by other small values shrink, until the RNN passes along values which are too small to impact subsequent computations. A technique which addresses this problem is the use of memory gates, most famously through **Long Short-Term Memory (LSTMs)** (Hochreiter and Schmidhuber, 1997). Memory gates constitute a refinement of the matrix U from Eq. 2.2.2.2 and introduce trainable selective updating of memory. This allows information of high importance to be held onto for longer if needed. This in turn makes models converge faster and perform better, especially at predicting labels for long sequences.

For even more powerful modeling, LSTMs can be instructed to process sequences in a bi-directional manner. One-directional models like that shown in Figure 2.2 only preserve the past and only learn dependencies between preceding and subsequent items in a sequence. **Bi-directional LSTMs (BiLSTMs)** like that shown in Figure 2.3 pass over sequences in a left-to-right and right-to-left manner in parallel. This produces the hidden states $\overrightarrow{\mathbf{h}_T}$ and $\overleftarrow{\mathbf{h}_T}$, which are then concatenated or pooled into a single hidden state. As a result, updates to the model based on token $\mathbf{x}_i$ take both its relationships with past and future tokens into consideration, resulting in better encoded representations $\mathbf{z}$. To use a BiLSTM layer for token-level classification, the hidden states $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$ corresponding to $\mathbf{x}_i$ are concatenated

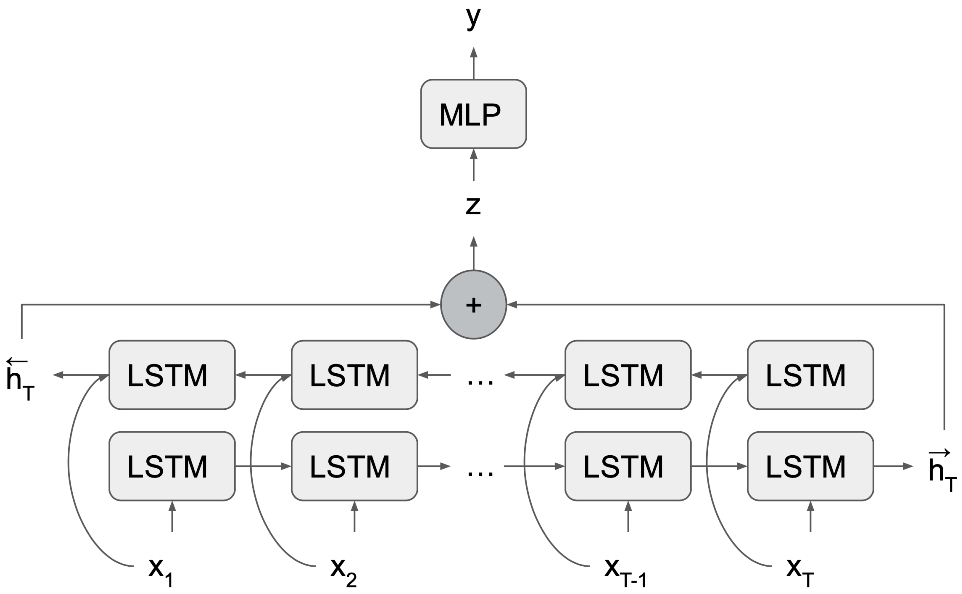

Fig. 2.3 Diagram illustrating a BiLSTM for sequence classification: the BiLSTM encoder makes a forward and a backward pass over the input sequence. The resulting hidden states $\overrightarrow{\mathbf{h}}_T$ and $\overleftarrow{\mathbf{h}}_T$ are concatenated to obtain a encoded representation **z** and transformed by an MLP into a label **y**.

.

or pooled and then directed to an Multi-layer Perceptron (MLP) which converts them into a token-level label $\mathbf{y}_i$. For use for sequence-level classification, the final hidden states of the pass ($\overrightarrow{\mathbf{h}}_T$ and $\overleftarrow{\mathbf{h}}_T$) are instead concatenated or pooled to sequence representation **z**, which is then processed by an MLP into label **y**. This use is illustrated in Figure 2.3

2.2.2.3 Transfer Learning and Pre-training

A drawback of deep learning methods is that as a result of the large number of parameters they have, they are prone to committing properties of the training dataset to memory in such detail that they remember relationships between features and labels that are a coincidence of the training data. This problem is called the problem of over-fitting. There exist various hyperparameters that can be tuned to help neural architectures avoid over-fitting and generalise better, but these on their own may not completely make up for a lack of data or data diversity. To solve this problem, it helps to visualise an NLP task as made up of two distinct problems. The first is the problem of modeling textual data made up of discrete symbols into a feature vector consisting of continuous values $\mathbf{x} \in \mathbb{R}^{d_{in}}$ that represents the meaning of these symbols accurately and in enough detail. The second is to train a classifier that learns task-specific assignments for these continuous token or sentence representations, which allows it to predict e.g. part-of-speech tags or sentiment labels.

A straight-forward, simple approach to create numerical representations of text is **one-hot encoding**. This method treats each unique token in the input data as a feature. Each input sentence is represented as a feature vector of the same size as the number of tokens in the input data, i.e. the same size as the vocabulary. The feature vector is made up of zeros and ones. A zero indicates that a token is not present in the input sentence and a one indicates that it is. This encoding method is not particularly efficient, because a lot of space is token up by each instance to represent the absence of the many tokens in the large vocabulary that are not present in that instance.

An improvement over the one-hot-encoding method is to use **word embeddings**. Word embeddings are the outcome of a training process on unlabeled data whereby co-occurrence patterns between words are embedded as vectors in a feature space. The notion that co-occurrence statistics are an effective way to represent meaning is derived from the distributional hypothesis which states that "a word is characterized by the company it keeps" (Firth, 1957). This concept of distributional semantics predicts that vectors of similar words will cluster together in their shared feature space. Because distributional word representations contain some redundancy, dimensionality reduction techniques can be applied to obtain representations that are not only much more fine-grained but also more efficient than one-hot-encodings.

To produce sophisticated word embeddings, language models are trained by taking one-hot encoded words as input, and training a neural network on the task of predicting the context of each input. Once the network has trained for long enough on a large and varied enough set of contexts, it will have learned for each input what context words it is most likely to co-occur with, and it will have stored that information in its weights. The first word embeddings were created with neural algorithms that produce fixed embeddings for words regardless of their context. The most popular pretrained non-contextualised word embeddings are Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The current state-of-the-art is to use contextualised **neural language models** such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) which capture the meaning of a word in different contexts.

To transfer the knowledge acquired from unlabeled data during the pre-training step, language model weights can be extracted and used as feature vectors for downstream tasks. Alternatively, the first layer of a language model can be used directly as the first layer of a classifier architecture. In the latter case, this initial layer is a matrix $\mathbf{E} \in \mathbb{R}^{|vocab| \times d}$ that maps all words in the vocabulary to vectors of dimension d. This layer can be frozen to prevent loss

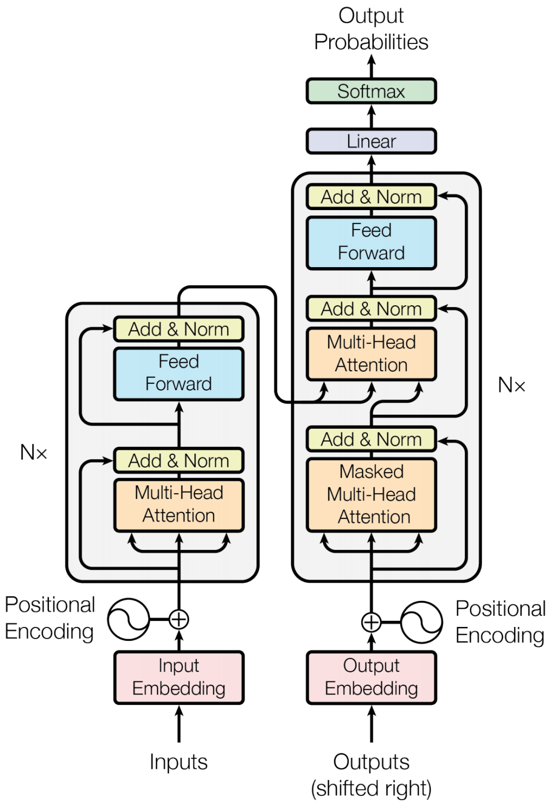

Fig. 2.4 The Transformer architecture as depicted in Vaswani et al. (2017)

of learnings from the pre-training phase, or it can be added to the learnable set of parameters that is adjusted during training on the downstream task.

The currently preferred architecture for neural language models is the **Transformer** architecture (Vaswani et al., 2017). This is a type of neural network that does not learn relationships between words in a sequence like LSTMs do, but instead uses an **attention mechanism** to selectively learn whatever relationships are most important between not only a word and previous words in the sequence, but the target word and any other words in the sequence.

Transformers consist of an embedding layer, an encoder and a decoder (see Figure 2.5). The encoder consists of a stack of layers with a multi-head self-attention mechanism and a fully connected feed-forward network. Residual connections are used after both the self-attention layers and the FFNN as well as Layer normalisation (Ba et al., 2016) is applied to both the multi-head attention component and the FFNN and incorporates residual connections (He et al., 2016) in each case as well. The multi-head self-attention mechanism within the encoder is a mechanism whereby several copies (heads) of an attention component are in place to allow the network to attend to different sections of the input sequence as needed.

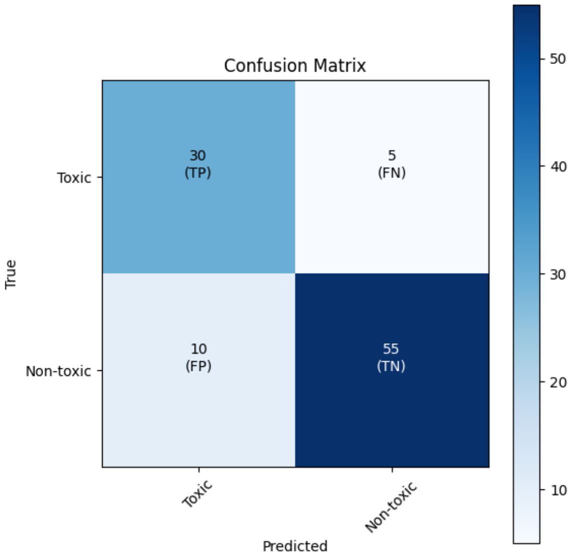

Fig. 2.5 An illustration of a confusion matrix that represents model performance at a toxic language classification task. It distinguishes between true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Note that this mechanism contains parameters that are to reflect relationships between each input token and any or even all of the other input tokens in the input sequence. It is this feature that makes the Transformer both more powerful but also costlier to train than LSTMs. The decoder consists of an N-sized stack of multi-head attention layers. The decoder furthermore contains a masked self-attention layer that only attends to previously decoded tokens, thus forcing predictions for position i to depend on the known outputs at positions less than i.

2.2.2.4 Evaluation

Supervised learning methods are evaluated by comparing predictions of a trained model on held-out test data with the gold labels for those instances. The most straight-forward metric is to compute what number of predictions was correct. This metric is called **accuracy**. In most scenarios, however, we want to describe model performance in more detail.

The most commonly used approach for performance evaluation that is more refined than measuring accuracy expresses model performance by categorising predictions according to a *confusion matrix*. A confusion matrix distinguishes between true positives, false positives, true negatives and false negatives. Using binary classification problem with the classes *toxic* and *non-toxic* as an example, a true positive (TP) is a correct identification of a toxic instance, a true negative (TN) is a correct labeling of an instance as non-toxic, a false positive (FP) is an incorrect flagging of an instance as toxic, and a false negative (FN) is a failure to identify a toxic instance.

The confusion matrix can be used to compute scores that reflect a trade-off between precision and recall. Using the same example, the **precision score** states how many flags of instances as toxic were correct. The precision score is defined as $P = \frac{TP}{FP+TP}$.

The **recall score** states how many toxic instances were caught by the system, and is defined as $R = \frac{TP}{FN+TP}$.

In some scenarios, the ideal model recognises as many instances of a class as possible, even if this causes the model to make frequent mistakes. In other use cases, the ideal model only labels instances as belonging to a certain class if it is very confident in its decision. In most cases, the preferred model behaviour is one which strikes a balance between precision and recall across its class. The most common metric for supervised models is therefore the **F1-score**: the harmonic mean between precision and recall, given by $F1 = 2 \times \frac{precision \times recall}{precision+recall}$."

In the case of binary classification problems, only a single F1-score needs to be reported. In the case of multi-class classification, it is common to report either the macro-average of the F1-scores, which treats the performance of all classes as equally important, or a weighted average of F1-scores that takes into consideration the size of each class.

Chapter 3

Related Work

3.1 Introduction

Studies on automatic detection and classification of media frames are found in NLP publications and in Communication Science and Political Science publications, where they are a part of text-as-data research. At the time of writing this thesis, work on automatic framing detection is a relatively new phenomenon in both NLP and text-as-data research. As such, commonly accepted definitions and terms are still being developed. To give an example, some NLP work that focuses on selective presentation of aspects of a topic does not call this framing, but instead uses other terms such as spin or implicit sentiment. We consider a study to be an example of what we will call Framing Detection (FD) regardless of the terminology practices in the paper, if its primary purpose is to detect or classify parts of a document that may influence readers' interpretation of an issue or entity through selectively mentioning or omitting aspects of that issue or entity.

In the second half of this chapter, we discuss work that resembles Framing Detection, but in which explicitly evaluative text is not excluded from the scope of what the system is meant to detect. An important characteristic of Framing Detection is that the positive or negative interpretation of the target issue or entity is arrived at through inference, and that there is no assumption of conscious awareness or intentional effort on the part of the writer that their wording has the potential to influence a reader's opinion regarding the issue or entity at hand. This sets FD apart from Bias Classification and Opinion Mining. **Bias classification** involves the categorisation of texts into different viewpoints on a topic (e.g. in favour of or against same-sex marriage) or different sides of sociopolitical spectrum (e.g. left-wing or right-wing). It aims to detect any phenomena that help identify what sociopolitical identity that text is aligned with or which side of an argument the author likely identifies with. The difference with Framing Detection is not only the inclusion of explicit indicators of opinion, but also a focus on document classification. Also included in the second half of the chapter is work on explicit sentiment (i.e. extraction and/or classification of explicit evaluations of targets), which we refer to as **Opinion Mining**. Following the literature, we consider Opinion Mining to involve the detection or classification of text which expresses sentiment, appraisals and feelings toward targets (Pang et al., 2008; Liu, 2012). The concept of opinion here is broad and can include both strongly and weakly subjective text (Cambria et al., 2017). Consequently, we cover a variety of tasks under this header, including subjectivity detection, sentiment classification, stance detection, abusive language detection and connotation classification.

Framing, bias and opinion are related concepts, and we recognise that the categorisation of related work used in this chapter is not the only reasonable categorisation. We also acknowledge that some studies qualify for classification under more than one of the proposed categories.

3.2 Framing Detection

3.2.1 Issue Framing

Most work on FD focuses on the framing of issues and topics, such as abortion, same-sex marriage, or climate change. The manual approach to systematically analyzing issue framing is to use a coding methodology. This methodology consists of selecting a set of frames to study, gathering a corpus of documents, and marking each section of text in the corpus that pertains to one of the selected frames according to a coding schema. Coding frames is labor-intensive work, requiring careful reading of the data and expertise on the part of the annotator. For this reason, issue framing research often uses unsupervised techniques to automatically detect framing-like patterns in unlabeled data, as described in Section 3.2.1.1. Other studies experiment with the automation of issue-specific frame coding with training supervised systems, discussed in Section 3.2.1.2.

3.2.1.1 Unsupervised Approaches

Unsupervised approaches include the use of clustering and LDA-based approaches to group together portions of text that can be regarded as an issue-specific way of framing a (sub)-topic.

Nguyen et al. (2013) and Nguyen (2015) apply various topic modeling techniques including an LDA-based approach to model a multi-level topic hierarchy from transcripts of interactions between U.S. legislators. The higher levels represent key issues (e.g. *macroeconomics* or *health*) while the lower levels of the hierarchy are considered to reflect framing modes (e.g. "disadvantages of government shutdown" as a framing mode for *macroeconomics*, or "obamacare as a government takeover" as a framing mode for *health*). This approach allows the authors to analyze which general topics are most polarizing between legislators, and subsequently which different framing modes are used by groups of legislators when debating each key topic. For example, when applied to congressional bill data from 2011-2013, the topic hierarchy suggests that on the key issue *macroeconomics* some Republican legislators focused on criticizing government overspending, while others focused on the downsides of a government shutdown.

Tsur et al. (2015) use probabilistic topic models (also LDA) with time series regression analysis to analyze development of framing strategies over time. Their proposed approach infers topics from public statements of U.S. congress members over a four-year period. Smaller topics, which are regarded as issue-specific frames, are aggregated into larger clusters which are regarded as topics.

Menini et al. (2017) take a different approach to unsupervised modeling of issue frames. Their method involves classifying manifestos into macro-domains and extracting key concepts by averaging the embeddings of words in candidate phrases. They show that modeling framing strategies through key concept extraction and clustering is more effective than either *k*-means clustering, a graph-based approach or standard LDA-based topic modeling for the downstream task of classifying agreement between party manifestos.

Demszky et al. (2019) also cluster embeddings rather than using an LDA-based approach to model framing patterns. Their unsupervised approach uncovers differences in how Republican and Democrat Twitter accounts discuss mass shootings on a dataset of 4.4M tweets on 21 U.S. mass shootings.

3.2.1.2 Supervised Approaches

To the best of our knowledge, Burscher et al. (2014) were the first to publish an attempt to apply supervised ML to the coding of frames. They explored automatic coding of four generic frames, but struggled to achieve reasonable performance due to data scarcity. In the following year, Baumer et al. (2015) published the results of training Naive Bayes classifiers on various feature sets aimed at capturing issue framing in U.S. news texts. They found that imagery, figurativenes, and lexical features were indicators of framing language, and found that their computational approach to issue framing detection reached close to human-level performance on their data set of 75 documents.

In a publication describing their collaboration Boydstun et al. (2014) proposes the use of the Policy Frames Codebook (Boydstun et al., 2014) to generate sufficient data to effectively develop computational approaches to framing annotation and detection. This proposal was followed by the release of the development and publication of the Media Frames Corpus (MFC) (Card et al., 2015), a substantial collection of U.S. news annotated for fifteen generic issue frames. The corpus consists of around 16k news articles on immigration, smoking and same-sex marriage. The news texts are annotated for fifteen framing dimensions, including for example the *morality* frame and the *economic* frame. The inter-annotator agreement on frames for the different frames ranged between 0.08 and 0.23 in terms of Krippendorff's α (Krippendorff, 2018).

The MFC precipitated the influx of attempts to solve the problem of automatic issue framing labeling with supervised approaches. Naderi and Hirst (2017) compared the performance of several neural methods on document-level classification on the MFC. Their best-performing model used 300-dimension GloVe word embbedings for input representation and a memory gate mechanism similar to that of LSTMs for classification, achieving an accuracy of 58.7% and an F1-score on of 57.1% for 15-way classification on a held-out portion of the MFC's immigration subset. On this same immigration subset, Ji and Smith (2017) and Khanehzar et al. (2019) achieved accuracy scores of 58.4% and 65.8% respectively. Ji and Smith (2017) approached the problem with a novel neural discourse structure network with an attention mechanism, while Khanehzar et al. (2019) used the pre-trained language model RoBERTa and also reported on a novel corpus based on Australian parliamentary speeches. Cabot et al. (2020) used the entire MFC corpus, and achieved an accuracy score of 71.6% with a joint model of metaphor, emotion and political rhetoric within a multi-task learning framework based on pre-trained RoBERTa.

Other work on the MFC includes Field et al. (2018) and Johnson et al. (2017). Field et al. (2018) used MFC data to create a lexicon for each one of the 15 frames. This lexicon was projected to 100K articles of Russian news with embedding-based methods, which allowed the authors to identify trends in reporting on events that concern the United States. Johnson et al. (2017) built upon the 15 issue frames presented in the MFC to analyze statements by politicians on Twitter. This work extends the number of frames to include additional frames that are particular to the social media domain. The authors also propose a feature-based probabilistic approach to predict the use of issue frames in tweets.

Another substantial corpus for issue framing research is the Gun Violence Frame Corpus (GVFC) (Liu et al., 2019a). This collection consists of around 3k headlines of articles from well-known U.S. news websites of which 1,300 pertain to gun violence. The data is annotated by experts for 9 frames: four generic frames (Politics", "Public opinion", "Society/Culture", and "Economic consequences") and five issue frames specific to gun violence coverage ("2nd Amendment", "Gun control/regulation", "Mental health", "School/Public space safety", and "Race/Ethnicity".

Liu et al. (2019a) test several approaches on this data set, including the approach developed by Field et al. (2018) for Russian news. Their best method outperforms Field et al. (2018), and is then used to study 88k unlabeled news headlines to analyze gun violence coverage on a larger scale. Akyürek et al. (2020) work on detecting issue frames in the GVFC in an exploration of multi-label multi-lingual task issue framing detection. They apply multilingual transfer learning along with the application of a dictionary to low-resource scenarios. Tourni et al. (2021) extend the GVFC with images and explore the effectiveness of a multi-modal approach.

Other supervised approaches to issue frames include a text-as-data study by Opperhuizen et al. (2019) and a study of online discussions by Hartmann et al. (2019). Opperhuizen et al. (2019) present and analyse a collection of 2265 Dutch newspaper articles for framing of earthquakes caused by gas drilling in the north of the Netherlands. Hartmann et al. (2019) introduce an issue-frame annotated corpus of online discussions, and use multi-task and adversarial training to apply issue frame detection models from the news and social media domain to this data. Mendelsohn et al. (2021) perform a computational analysis of a dataset of tweets on immigration labeled for Policy Frames Codebook frames (Boydstun et al., 2013) as well as some immigration-specific frames, using amongst other techniques, a fine-tuned RoBERTa model.

3.2.2 Entity Framing

One thing to be noted about work on Entity Framing is that it is very scarce. The limited number of papers that exists can be divided into three categories. The first category consists of work on the modeling of latent characterisations of entities that resembles the use of LDA models in issue framing analysis. The second category contains work which analyses implicit subjectivity through syntactic choices. The third category, which is the most relevant to the experiments presented in this thesis, consists of studies of patterns in and possible impact of the inclusion of objectively verifiable properties when referring to entities.

3.2.2.1 Latent personas

Motivated by theoretical work on recurring representations of characters in the news (Schneider and Ingram, 1993; Van Gorp, 2010) (see also Section 2.1.3.3), Card et al. (2016) propose an adaptation of Bamman et al. (2013)'s Dirichlet Persona Model (DPM) to model latent characterisations of entities in news. The original model, an unsupervised graph-based model developed to analyse named characters in movie narratives, is extended to apply to unnamed non-person entities and to perform co-occurrence-based clustering.

The authors apply an extended DPM to MFC articles and identify recurring characterisations of unnamed entities as e.g. *refugees*, *workers* and political *candidates*. Their method also obtains casts, which are clusters that specify mention words associated with personas as well as co-occurrence information from relation tuples containing the personas. Finally, the authors experiment with the use of personas and casts as features for predicting the primary frame and tone of articles, and find that they are predictive of a text's primary frame, but not more effective than existing features such as semantic frames.

3.2.2.2 Syntactic Structures

Work on framing through syntactic structures investigates how grammatically relevant properties can give rise to inferences that may influence opinions (Greene, 2007). A key example is framing through transitivity, where roles are filled or left unfilled in ways that can impact impressions of culpability. For example, Ronald Reagan famously stated that "Mistakes were made" in the Iran-contra affair (Broder, 2007). Using the passive form here allowed Reagan to admit fault in the abstract, without assigning responsibility to an identifiable person or institution.

Greene and Resnik (2009) identify 13 such grammatically relevant properties of events and participants that are capable of conveying implicit sentiment. These features describe aspects of meaning of elements in the canonical transitive clause *X verb Y*, such as *volition*, *telicity*, and *affectedness*. They can be divided into features related to semantic transitivity (from Dowty (1991)) and features related to transference of verbs to arguments (from Hopper and Thompson (1980)). Formal transference of meanings from verbs to arguments has been shown to reflect the inferences humans make about participants in events based on the semantics of the verb (Kako, 2006).

Greene and Resnik (2009) propose a method of identifying values for these 13 dimensions by means of so-called *observable proxies for underlying semantics* (OPUS), as a stand-in for directly observing the 13 dimensions of meaning. OPUS features are created by first identifying domain-relevant terms by comparing the target domain to the British National Corpus (Leech et al., 1992), and then specifying for each training instance whether these relevant terms appear as verbs or arguments, whether the verbs are used transitively or not, and what role the arguments fulfill as indicated by the dependency parse of that instance. The OPUS feature extraction method is used to train SVM and Naive Bayes models on a novel corpus of pro- and anti-death-penalty documents, as well as on a corpus of Israeli and Palestinian viewpoints on prominent issues of concern (Lin et al., 2006). Experimental results show that human-annotated sentiment ratings can be predicted by models trained on values for the set of OPUS features with higher accuracy than competitive classification systems using other features.

3.2.2.3 Objectively Verifiable Attributes

In contrast to the Entity Framing work on latent personas and syntactic structures, the papers described in this section all directly analyse how choices writers make about including or omitting mentions of entities' objectively verifiable attributes may impact readers' impressions of these entities.

Family relationships. Wagner et al. (2015) apply a computational method to examine differences in the portrayal of men and women across multiple dimensions in six language editions of Wikipedia. They find equal coverage of men and women in all editions in terms of frequency and visibility, which they consider evidence that an effort is being and/or has been made to mitigate gender bias on the platform. However, they find that there are more links connecting women to men's pages than vice versa. Additionally, they find a high number of terms describing family relationships, such as *married* and *mother*, in the list of terms that are more frequently mentioned in articles about women than men. This may be interpreted as evidence that framing through the mentioning of objectively verifiable attributes like martial status or parenthood is a subtle phenomenon that can crop up as an unwanted difference in portrayals of entities of opposing sexes, even where authors have made efforts to avoid such differences. Follow-on work (Wagner et al., 2016) confirms the finding that the biographies of women more frequently mention family relationships and romantic relationships. This work also shares other findings that are suggestive of gender bias (i.e. biographies of women cover figures of higher notability than biographies of men, and abstract terms describe negative aspects when used in biographies of women and positive aspects when used in biographies of men) (Wagner et al., 2016).

Attributes of police shooting victims. Ziems and Yang (2021) create and analyze a corpus of 82k U.S. news articles on instances of police violence with a fatal ending: the Police Violence Frame Corpus. They computationally parse attributes of victims to compare the way liberal and conservative media portray them. The authors find that liberal news coverage tends to emphasize the victim's race, and are more likely to mention that victims were unarmed. Conservative sources are more likely to mention that victims had a criminal record, and that they were armed. The work also explores other attributes and aspects that are more subjective (e.g. whether victims were aggressive) or not entity-centric (e.g. whether the killing was an example of social injustice).

Background information. Fan et al. (2019) present the first and only corpus labeled for informational entity framing, as well as the first supervised approach to the detection of such framing. The authors are interested in IEF, which they refer to as ‘ìnformational bias”, as well as in lexical bias. Lexical bias is defined as bias that can be mitigated through rephrasing, while informational bias has the familiar definition of bias through the inclusion of tangential or background information that can colour a reader’s perception of an entity.

Fan et al. (2019) present a novel data set called the BASIL (Bias Annotation Spans on the Informational Level) corpus, which consists of 100 triples of U.S. news articles from Fox News, the New York Times and the Huffington post. Each article was given a variety of document-level, sentence-level and span-level annotations, including annotations for lexical bias spans and IEF spans along with their polarity and target entities. Because the BASIL corpus is the training data for the experiments presented in Chapter 5, Section 5.2 gives a more detailed description of BASIL’s contents and the annotation procedure used to obtain labels.

In addition to presenting the BASIL corpus, Fan et al. (2019) train straight-forward BERT-based classifiers for both token-level and sentence-level classification of lexical bias and IEF. Although they mention that human annotators read sentences in their context to provide gold labels, the presented automatic classification approach to IEF models and labels sentences in isolation. Performance scores are higher for lexical bias detection than for IEF detection, illustrating the higher difficulty of the second problem.

Names and titles. As mentioned in Section 2.1.3.3 on Entity Framing Theory, writers can use forms of reference to convey not only referential but also indexical meaning, the latter of which carries implicit sentiment towards entities (Richardson, 2006; Blommaert et al., 2005). Some of this implicit sentiment can be linguistically encoded, for example in the politeness of the chosen form of reference. The possible framing effects of forms of reference, including of names and titles, have not appeared on the radar of text-as-data or NLP research, but have been studied implicitly to quite some extent by sociolinguists, whose work forms the theoretical foundation of the naming and titling experiments described in Chapter 4.

According to sociolinguists Brown and Gilman, the semantics of reference and address should be broken down into significations of differences in power and of difference in solidarity (Brown and Gilman, 1960). When languages do not possess separate markers for perceived levels of superiority or inferiority and for perceived levels of solidarity and familiarity, like is the case for English, address signifiers may overlap in meaning and therefore be ambiguous.

Ervin-Tripp (1972) identifies various other factors that play into naming form usage including a) personal characteristics of the addressee (age, kin, sex, marital status), b) situational characteristics (formality of the context, location of the interaction), c) cognitive characteristics (e.g. whether the name is known) and d) cultural characteristics (this includes roles and titles unique to a culture). Choosing the idiomatic term of address in a certain situation is a type of linguistic knowledge, which takes into account characteristics of individuals such as their sex, as well as contextually relevant aspects of relationships between individuals such as their position in a professional hierarchy.

3.3 Subjectivity Analysis and Detection

The detection of framing in NLP constitutes a small field compared to the much larger body of related work on subjective language of various kinds, including partisan language, lexical bias, evaluative language, abusive language, and connotation. In this section, we cover work on partisan language and lexical bias as Bias Classification in Section 3.3.1, and we cover work on evaluative, abusive, or connotation-bearing language as Opinion Mining in Section 3.3.2.

3.3.1 Bias Classification

We define Bias Classification (BC) as work that attempts to detect differences in writing style that are caused by differences in opinion and world-view. Even when they are not making their opinion on the topic explicitly known, individuals may "give away" which sociopolitical group they belong to and consequently which stance they are likely to have on a topic by e.g. using different examples or by calling things by different names (e.g. "pro-life" rather than "anti-abortion"). Other terms used to refer to what we consider to be the same phenomenon include *perspective*, *point-of-view*, *slant*, *ideology*, *orientation*, *leaning*, or *partisan language*.

Text-level Bias Classification Most bias labeling systems provide text-level predictions, i.e. they state for a single document what side the author of the article most likely identifies with. Lin et al. (2006) approach this problem by using statistical supervised learning to identify sentences in news articles that are strongly indicative of bias. Their method outperforms a simple baseline on the Bitter Lemons data set of articles about the Israeli-Palestinian conflict. Fulgoni et al. (2016) use handcrafted lexicons that reflect differences in moral values. This approach is based on insights from Moral Foundations Theory about systematic differences in how conservative and progressive people view the world (Haidt and Graham, 2007; Haidt and Joseph, 2004). Their handcrafted lexicons help predict partisanship of articles on 17 controversial topics including abortion and climate change. Kulkarni et al. (2018) show that bias detection benefits from adding network information as a feature along with textual information. Other work targets text genres other than news including online discussions and Wikipedia pages (Iyyer et al., 2014; Kato et al., 2008; Yano et al., 2010; Johnson and Goldwasser, 2018; Recasens et al., 2013).

Some text-level BC work focuses specifically on lexical bias, i.e. bias revealed through journalists' word choices (Hamborg et al., 2019a,b). Annotations for lexical bias are provided by the BASIL dataset (Fan et al., 2019) and the NewsWCL50 dataset (Hamborg et al., 2019a).

A final sub-area of text-level BC is the manipulation of the bias of a text with generative methods. Chen et al. (2018) experiment with this by transforming biased article (either *left* or *right*) into an article with the same topic but the opposite bias label using an autoencoder.

Outlet-level Bias Classification Outlet-level BC systems predict bias labels or ratings for sources of articles rather than individual articles. Gentzkow and Shapiro (2010) index U.S. news outlets by comparing their language to that of a congressional Republican or Democrat. Groseclose and Milyo (2005) base their predictions on articles' citing behaviour. Baly et al. (2018) demonstrate that outlet-level BC, like text-level BC, benefits from including features that encode things other than textual information, in this case urls, Twitter data, the outlet's Wikipedia page, and information about traffic to the outlet's website.

Some outlet-level BC specifically targets news outlets or articles that fall on extreme ends of the political spectrum. This task, known as Hyperpartisan News Detection, is used as a prior step to the task of fact-checking (Kiesel et al., 2019; Jiang et al., 2019; Potthast et al., 2018).

3.3.2 Opinion Mining

The detection of framing in NLP builds upon extensive literature in Opinion Mining and Sentiment Analysis (Pang et al., 2008; Liu, 2012). In this literature, opinion is considered to be text which expresses sentiment, appraisals or feelings toward targets. This concept of opinion is broad and can include both very strongly and very weakly subjective text (Cambria et al., 2017). Opinion Mining thus encompasses a number of NLP tasks, including the canonical tasks Subjectivity Detection, Sentiment Classification, Stance Detection and Abusive language detection, and the niche task sof connotation modeling and opinion citation analysis.

Canonical Tasks. Subjectivity Detection is the task of separating objective from subjective text (Pang et al., 2008; Wilson et al., 2005). It removes factual content from text as a pre-processing step prior to Sentiment Classification. Sentiment Classification is the task of determining the polarity of a span of subjective text (Lin et al., 2011; Wiebe et al., 2004). Aspect-based Sentiment Classification detects which aspects of target of evaluation the subjective text comments on (Do et al., 2019; Pontiki et al., 2016). Stance Detection is a more fine-grained task, and involves classifying opinions along with their targets, and in some cases along with the relevant aspects of targets (Mohammad et al., 2017; Taddy, 2013; Pontiki et al., 2016). Abusive Language Detection, sometimes referred to as Hate Speech Detection, aims to detect text that expresses opinions and feelings that are so negative, offensive or aggressive, that they warrant removal from public platforms like Twitter or Facebook (Davidson et al., 2017; Zampieri et al., 2019, 2020).

Connotation Frame Analysis. Connotation frames draw inspiration from frame semantics (Fillmore et al., 1976), but capture connotational rather than denotational meaning (Rashkin et al., 2015). The connotation frame model consists of 5 elements: (Rashkin et al., 2015, p.1)):

1. *Writer's perspective*: what attitude the predicate suggests the author has towards each participant.
2. *Entities' perspective:* what attitude the predicate suggests the participants have towards each other
3. *Effect:* what effect (positive or negative) the event has on the participants.
4. *Value:* what the predicate suggests about the value of the participants.
5. *Mental state:* what mental states the participants can be assumed to have as a consequence of the event described by the predicate.

Connotation frames were extended from English to other European languages using a collection of Twitter data. As a result of this extension, there are now 1.2 million connotation frames available in 11 languages (Rashkin et al., 2017). Connotation frames have also been extended to include power and agency dimensions, which has been used to conduct studies on gender bias in films (Sap et al., 2017) and in coverage of the #MeToo movement (Field et al., 2019).

Opinion Citation Analysis. Opinion citation is a strategy for introducing evaluative language while technically maintaining neutrality. Although publications on the topic are rare, it has been shown to be a common tool for framing in news texts (Fan et al., 2019), Niculae et al. (2015) quantitatively measure bias in quoting patterns in a data set of U.S. political news and presidential speeches published during Barack Obama's presidency by encoding quoting patterns in a low-rank space. They show that quoting patterns are indeed indicative of bias that indicates sociopolitical division. Luo et al. (2020) analyze linguistic devices for citing third-party opinion citation in a novel stance-labeled data set of climate change discourse. They find that climate change sceptics tend to quote opponents using language that shows distrust of the cited opinions, and that both sides of the debate exhibit a pattern of citing third-party opinions that are at odds with their known stance regarding climate change.

Chapter 4

Framing Through Naming and Titling

4.1 Introduction

The most elementary choice one makes when discussing a political entity is what form of their name to use when referring to them. In news articles, the names and titles which may be used for entities are carefully prescribed by style guides (Siegal and Connolly, 1999; Raue, 2012). These rules for name and title usage suggest that names and titles convey subjective meaning. This is also suggested by the complaints newspapers sometimes receive from readers that their usage of politicians' names is not sufficiently respectful[1]. In this chapter, we quantify that names and titles convey subjective meaning conveyed by demonstrating a statistically significant relation between the names and titles used for entities on the one hand, and writers' perceived attitude towards those entities on the other hand.

As discussed in Section 2.1.4, there is evidence from Communication Science studies that framing in text influences attitudes towards policies and topics. But there are no studies on the specific influence of naming and titling on attitudes towards entities. We therefore base our hypotheses instead on sociolinguistic work. Sociolinguists have observed two recurring subjective functions of naming: marking differences in status (Brown and Gilman, 1960; Brown and Ford, 1961; Dickey, 1997), and marking differences in degree of membership to the same group (Allerton, 1996; Brown and Gilman, 1960; Brown and Ford, 1961). These observations were made during small-scale studies of spoken language, where names are used either as a form of address to refer to a conversation partner or as a form of reference to a third party. In English and related Germanic languages, both the signalling of status as of solidarity is done through variation in formality (Dickey, 1997; Brown and Gilman, 1960; Brown and Ford, 1961; Allerton, 1996).

[1] https://www.nytimes.com/2017/11/08/reader-center/why-does-nyt-call-president-mr-trump.html

Based on sociolinguists' findings regarding the function of naming, we predict that there are two possible answers to the first research question of this thesis, which asks whether there is a measurable relation between naming/titling and stance in social media messages that mention politicians. Given the dual function of naming, our hypotheses regarding the relation between naming variation and stance are: H1) Naming formality primarily signals status, and is therefore positively correlated with author stance, and H2) naming formality primarily signals solidarity, and is therefore negatively correlated with author stance. Note that these hypotheses do not touch on the topic of how readers adjust their opinion in response to the connotation they assign to the naming/titling choice. Changes in attitude are a psychological phenomenon that depends not only on word-choice but also on subject-dependent factors. For example, a reader's opinion of an entity may actually solidify when they perceive a writer's attitude to be the opposite of theirs (Slothuus and De Vreese, 2010) (see also Section 2.1.2.1). Rather, we demonstrate quantitatively, on a large and varied sample of data, that a measurable association exists between *naming/titling*, an objective property of a person, and the perceived non-objective attitude or *stance* of the *writer* towards that entity.

In addition to claiming status-indicating and solidarity-indicating functions of naming, sociolinguists have also observed that the meaning of names and titles is context-dependent (Dickey, 1997). To account for this context-dependence, we examine two factors which we suspect influence the relation between naming/titling and stance.

The first is **political leaning**. Research in moral psychology suggests that sociopolitical communities differ in how much value they place on establishing and maintaining hierarchies (Haidt and Joseph, 2004). This implies that the direction and strength of the relation between naming formality and stance may not be identical for writers' from different sociopolitical groups. Concretely, we suspect a difference between the strength of the positive connotation of formal naming in left-leaning and right-leaning discourse and aim to test this hypothesis on our data. To test this prediction, we investigate differences between posts by progressive and conservative social media users.

Another factor that we expect to be important is target entities' **gender**. There is some evidence that female politicians are referred to less formally than their male counterparts (Uscinski and Goren, 2011), as well as strong evidence that professional titles are used more frequently for male than female professors (Takiff et al., 2001), . We test whether a similar gender bias exists in the application of titles in a large-scale data set of Twitter mentions of male and female politicians in comparable political roles.

Section 4.2.1 and Section 4.2.2 describe the collection and annotation of our English-language and German-language Twitter corpora. Section 4.3 reports findings on the relation between naming variation and title usage. Here we establish the relation between stance and naming/titling in English-language mentions of presidents, and in German-language mentions of prominent parliamentarians who possess a doctoral title. Section 4.4 examines the impact of political leaning of the source and gender of the target entity on the relation between naming/titling and stance.

4.2 Corpus Construction

In this section we describe the construction of the two data sources used for our assessments of the relation between naming/titling and stance. The first is the English Twitter Titling Corpus (ETTC), a corpus of 4002 English-language tweets containing mentions of Presidents of large nations including or excluding the academic title *President*. The second is the German Twitter Titling Corpus (GTTC), a corpus of 1904 German tweets mentioning German parliament and government members with a doctoral degree by varying naming forms including or excluding the academic title *Dr.*. Both the ETTC[2] and GTTC[3] are publicly availablein a redacted form, where labels are associated with tweet IDs rather than tweet texts, as prescribed by the terms of use of the the Twitter API which aim to protect the right of users to remove their posts from Twitter at any time.

Both corpora consist of Twitter data because Twitter is a popular platform for political discourse (Tumasjan et al., 2010), where writers are not bound like journalists are to prescribed conventions for name and title usage (Siegal and Connolly, 1999). Tweets' short average length and frequent ungrammaticality usually limits their usefulness for computational sociopolitical research, but this is not an obstacle for our purposes, which require only discernible reference and stance.

We chose to construct our first corpus in English because of the ubiquity of the English language on social media platforms. The high volume of tweets in English maximises the likelihood of finding tweets which mention a diverse set of Presidents by a diverse set of naming forms, and thus the generalisability of our findings.

German is an interesting language to contrast with English because of stereotypes surrounding its relation with formality. It has been suggested that German speakers use titles more frequently than English speakers, but there are also indications that this is an outdated stereotype (Besch, 1998). There also exists literature which suggests that German speakers' political leaning impacts the formality of their terms of address. Left-wing student protests in the 1960s in Germany went hand-in-hand with a destabilization of conventions regarding formality in terms of address, including the use of academic titles (Besch, 1998). Instead

[2] https://doi.org/10.11588/data/IOHXDF
[3] https://doi.org/10.11588/data/IOHXDF

of using the polite pronoun *Sie* as the default among strangers, an increasing number of speakers treat informal *Du* as the default, and this convention has been associated with the political left (Glück and Sauer, 1997; Clyne et al., 2006; Besch, 1998).

4.2.1 Collection

4.2.1.1 English Twitter Titling Corpus

To obtain a collection of tweets mentioning a variety of political figures, we selected leaders who governed G20 countries with a presidential system in 2017 and who could therefore be referred to on Twitter with the professional title *President*. We also required that their names followed the order *first-name last-name*. This restriction avoids noise in the data due to confusion of the names of individuals from countries where the standard order of names is *first-name last-name*, like President Xi Jinping of China. Manual inspection of an initial collection of 50 tweets per country subset revealed that the Brazil subset consisted of very homogeneous tweets, and two others (Mexico and Argentina) contained many tweets that did not refer to the intended target. These subsets were therefore omitted from the data. The six political figures remaining after this selection process were: President Emmanuel Macron of France, President Joko Widodo of Indonesia, President Vladimir Putin of Russia, President Jacob Zuma of South-Africa, President Recep Tayyip Erdoğan of Turkey, and President Donald Trump of the United States of America.

We collected tweets from the Twitter API between 18 June 2017 and 30 August 2017 using three query types: LAST-NAME, #FIRST-NAME, and FIRST-NAME + (LAST-NAME/COUNTRY) and excluding retweets. As an example, the queries for France that were submitted to the case-insensitive API were *macron*, *#emmanuel*, and *emmanuel AND (macron OR france)*. We removed duplicates from the response, and reduced the number of headlines in the data, because naming variation in headlines is impacted by journalistic style conventions (Siegal and Connolly, 1999). We reduced the number of headlines by removing news tweets, which we defined as any tweet from an account with the string *news* in the username or profile description. The remaining tweets were then automatically labeled for their naming forms. Possible labels were: **first name only (FN)**, **last name only (LN)**, **full name (FNLN)**, **title and full name (TFNLN)** and **title and last name (TLN)**. We then sampled 1000 tweets per country, oversampling rarer naming forms to ensure a large enough number of tweets in each naming form groups. These tweets were then presented to annotators for relevance filtering.

To remove tweets that refer to a namesake rather than the intended target, we crowd-sourced on-target/off-target judgments by having tweets labeled as **on-target** or **off-target**.

Country Subset	**FE Agreement**	**Expert Agreement**
France	0.77	0.78
Indonesia	0.80	0.91
Russia	0.77	0.72
South Africa	0.77	0.87
Turkey	0.44	0.65
United States	0.65	0.78

Table 4.1 Inter-annotator agreement for the on-target/off-target task (Krippendorff alpha) between the three FE workers and between the two experts who adjudicated tweets where FE worker judgment was not unanimous.

Country Subset	**#Adj**	**Diff. Expert 1**	**Diff. Expert 2**
France	281	0.07%	0.05%
Indonesia	290	0.04%	0.03%
Russia	121	0.06%	0.06%
South-Africa	227	0.04%	0.04%
Turkey	128	0.05%	0.04%
United States	192	0.07%	0.06%

Table 4.2 Adjudication for the on-target/off-target task of ETTC tweets where FE worker judgment was not unanimous. #Adj provides the number of adjudicated tweets. Diff. Expert 1 and 2 give the percentage of tweets where the expert label differed from the FE majority vote.

Workers made their judgments on the platform Figure Eight (FE)[4] based on the prompt “Does this tweet mention President X?”. Possible answers were *yes*, *no*, *can’t tell* and *the link doesn’t work*. We collected 3 such relevance judgments per tweet. We adjudicated all instances where workers could not unanimously agree whether the tweet was on-target by supplementing the crowd-sourced labels with two additional judgments, provided by the thesis author and a research assistant. Table 4.1 shows the agreement between FE annotators and between expert annotators. We measured agreement with Krippendorff’s alpha (Krippendorff, 2018), which is suitable for multi-coder ordinal annotation (Antoine et al., 2014). We then compared expert relevance judgments to the majority vote from the FE relevance annotations and found very few differences (Table 4.2), suggesting that the majority vote is a reliable aggregation method for this annotation task.

All tweets which the majority vote labeled *no* (i.e. not relevant), *can’t tell* or *the link doesn’t work* were removed from the data set. Any which had been deleted from the platform

[4] https://www.figure-eight.com

since the relevance judgment were removed as well. On this smaller number of tweets we performed a more thorough deduplication method using a similarity threshold to discard duplicate tweets. Tweets which contained the president's name only in a trailing hash-tag were also removed.

Collected tweets which passed the relevance filtering phase were moved on to the stance annotation phase, during which they could still be discarded if a majority of annotators deemed the tweet impossible to label for stance. For this reason, we first describe collection of the GTTC in the next section, Section 4.2.1.2, and then the stance annotation procedure which applies to both the ETTC and GTTC in Section 4.2.2, before giving the final size statistics of both corpora in Section 4.2.3.

4.2.1.2 German Twitter Titling Corpus

To contrast findings on the ETTC with findings for another language, entity set, and type of title, we additionally constructed the GTTC from German tweets from June and July of 2018 mentioning prominent German parliament and government members who have a doctoral degree and can thus be referred to with the academic title *Dr.* To meet the prominence criteria, a politician must be or have been party chairman, Federal President, president of the parliament or member of the parliament after 2013. Out of the members of parliament with no additional function, we selected the top ten with the largest Twitter following. The list of politicians who passed these prominence criteria includes all politicians in Table 4.15, as well as two politicians (Barbara Hendricks and Johanna Wank) who later failed a title-containing tweet sampling criteria described near the end of this section.

We constructed queries with political keyword or party name disambiguators to increase the likelihood of receiving on-target results. The party names were: AfD, CDU, Bündnis 90/Die Grünen (henceforth B90/Grün), Die Linke, SPD and CSU. Table 4.4 provides the position of these parties on the political spectrum. Note that the German party FDP is not represented, because there were no FDP members who both had a doctoral degree and met our criteria for prominence after 2013. Two native speakers of German selected political keyword disambiguators to further improve the queries. These were: *Bundestag* (parliament), *MdB* (Mitglied des Deutschen Bundestages i.e. member of the German parliament), *Kanzlerin* (Chancelor), *Bundestagspräsident* (President of the parliament), *Regierung* (government), *Kabinett* (cabinet), *Ministerium* (ministry) and *Minister* (minister).

We did not create queries for mentions of politicians by their first name, as it would be exceedingly time-consuming to find relevant tweets among irrelevant ones mentioning common first names like "Thomas" or "Peter". In addition, first name references to politicians were found to be rare in English tweets mentioning presidents, and we expected naming by

Party	**Politician**	**#Tweets**
Die Linke	Sahra Wagenknecht	33
	Gregor Gysi	11
	Dietmar Bartsch	6
Subtotal		50
B90/Grün	Robert Habeck	23
	Anton Hofreiter	10
	Simone Peter	3
Subtotal		36
SPD	Franziska Giffey	103
	Karl Lauterbach	29
	Frank-Walter Steinmeier	20
	Karamba Diaby	15
	Katarina Barley	7
Subtotal		174
CDU/CSU	Angela Merkel	649
	Wolfgang Schäuble	32
	Gerd Müller	22
	Norbert Lammert	13
	Ursula von der Leyen	8
	Helge Braun	6
	Peter Tauber	6
	Thomas de Maizière	4
	Kristina Schröder	3
Subtotal		743
AfD	Alice Weidel	586
	Jörg Meuthen	193
	Alexander Gauland	103
	Frauke Petry	19
Subtotal		901
Total		1904

Table 4.3 Number of tweets per politician per party in the GTTC.

Party	Name	Political Position
Die Linke	The Left	Left-wing
B90/Grün	Alliance 90/The Greens	Centre-left/Left-wing
SPD	Social Democratic Party	Centre-left
CDU/CSU	Christian Democratic Union / Christian Social Union	Centre-right
AfD	Alternative for Germany	Right-wing

Table 4.4 Name and position on the political spectrum of the German political parties represented in the GTTC.

just the first name to be even more uncommon in the GTTC setting. There are some very prominent politicians that are occasionally addressed by their first name or a nickname, e.g. *Donald* for Donald Trump and *Vlad* for Vladimir Putin. Few German politicians have the degree of prominence required for this to be the case, meaning any findings of first name usage would generalise poorly.

The four types of queries were:

- #FIRST-NAME-LAST-NAME (e.g. #AngelaMerkel)
- FIRST-NAME LAST-NAME (e.g. Angela Merkel)
- LAST-NAME PARTY (e.g. Merkel CDU)
- LAST-NAME POLITICAL-KEYWORD (e.g. Merkel Bundestag)

We collected all tweets (excluding retweets) that matched one of these four query types over 6 weeks in June 2018 and July 2018. Results were limited to German language tweets as identified by a language detection algorithm. Journalists are instructed not to use the Dr. title unless topically relevant (Raue, 2012), so we excluded news tweets. For the GTTC, we defined as a *news tweets* any tweets from a manually compiled list of 247 news accounts (see Apppendix A). We assigned tweets a naming form type using a categorisation that is tailored to the purposes of our experiments. Unlike English, German allows titles to be combined, given rise to German equivalents of forms like *Professor Dr. John Smith* and *Mr. Dr. Smith*. This property makes the list of possible naming forms for German larger than optimal for meaningful comparisons between large enough groups of tweets. To address this, we annotated tweets as follows. In each instance, we first looked for the occurrence of a doctoral title followed by the target politician's first name and last name. If a match was found, the tweet was given the label **TFNLN**. If no match was found, we looked for the doctoral title and last name and assigned the label **TLN** if a match was found, otherwise we looked for first name and last name (labeled **FNLN**), and finally for the last name only (**LN**). This strategy automatically groups title/name combinations other than TFNLN, TLN, FNLN and LN in one of those four categories. For example, it places *Prof. Dr. John Smith* in the TFNLN category and *Frau Merkel* (Ms. Merkel) in the LN category. We discuss forms other than TFNLN, TLN, FNLN and LN and their relation to stance in Section 4.3.2.1. There were two politicians of the 26 who passed the criteria for prominence for whom no TLN or TFNLN tweets were returned by our querying process. Tweets for these politicians were removed, leaving 24 politicians with representation in the sample, listed in Table 4.15.

To ensure that the title-containing categories TFNLN and TLN are represented by enough tweets for a study on the impact of titling, we oversampled from these categories. We first sampled all TLN and TFNLN tweets. We then sampled the same number of LN and FNLN tweets in a stratified manner. To ensure no politician was represented exclusively by title-containing tweets, we applied on top of stratification the additional rule that each politician subset should contribute at least as many non-title tweets as title tweets to the overall sample. Because the queries used to sample tweets for the GTTC were more specific than those used for ETTC collection, there was no need for an on-target/off-target labeling step. Instead, we allowed annotators to label irrelevant or unreadable tweets as such during stance annotation (see Section 4.2.2.2).

4.2.2 Stance annotation

All the tweets sampled for the ETTC and GTTC were annotated for their stance towards the target politician on a 3-class stance scale. Because stance annotation requires more thorough comprehension of the text than on-target/off-target annotation, we switched crowd-sourcing platforms from Figure Eight to Amazon Mechanical Turk (AMT)[5], which gives the requester more control over the annotator screening process. AMT annotators' compensation was \$0.02 per HIT for approximately 7 HITs per minute. We asked annotators to respond to the stance annotation prompt described in Section 4.2.2.2, the possible answers to which are given in Table 4.5.

4.2.2.1 Annotator Screening

Workers who were interested in contributing labels were required to pass a language proficiency test and an instruction comprehension test. To protect against spammers, we only accepted annotators with a minimum number of 500 completed HITs and a minimum HIT approval rate of 97%. We flagged and removed annotators who did not maintain a task-internal

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

accuracy rate of 97% while annotating. The task-internal accuracy rate was based on trap questions making up roughly 4% of the data. For trap questions, we selected and annotated a subset of explicitly stance-expressing tweets. To give an example, one GTTC trap question contained a paraphrase of the text: "Dr. Wagenknecht gave a great speech today! Let it be known that there are still politicians with hearts and minds in Germany!”. We collected seven annotations per tweet.

4.2.2.2 Elicitation

The stance annotation elicitation prompt was based on the reader-perspective emotion label elicitation prompt in Buechel and Hahn (2017) which elicits emotion labels by asking: *How do you feel after reading this text.* We expect a reader-perspective prompt to better capture differences between tweets which are completely neutral in tone but reflect differently on the target, such as “Trump is trailing in the primaries” vs “The job market is improving under Trump”. Crucially, the prompt allows annotators to give a different rating to “President Trump visits France” than to “Trump visits France”. A disadvantage of reader-based prompts is their lower reliability (Buechel and Hahn, 2017). To combat this, our stance annotation elicitation prompt deviates from Buechel and Hahn (2017) by anchoring the perspective to that of a proponent of the stance target.

For the English annotators, the resulting prompt was: *How would a supporter of President X feel about this tweet?* The German prompt was: *Wie würde sich ein Anhänger von X fühlen, nachdem er/sie diesen Tweet gelesen hat?* (*How would a supporter of X feel after reading this tweet?*) Table 4.5 shows the possible responses.

Joseph et al. (2017) show that the quality of political stance annotations of Twitter data suffers when too little context is given during elicitation. We therefore provided annotators with the tweet location, user photo, user name and user description. If the tweet was a response to another tweet, that other tweet was shown also. Annotators were instructed to use the provided context as support when labeling ambiguous tweets.

ETTC	**GTTC**
positive	positiv
neither positive nor negative	weder positiv noch negativ
negative	negativ
cannot read / does not apply	nicht lesbar / trifft nicht z

Table 4.5 Possible responses to the stance annotation prompt.

4.2.2.3 Annotator Pruning with MACE

Dissatisfied with initial agreement scores, we used Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013) to identify and remove unreliable workers before taking a majority vote. Prior to applying MACE, we conducted an experiment akin to Hovy et al. (2013, p.2-3) which simulates adversarial conditions using synthetic data that resembles the ETTC and GTTC. We synthesised a set of 100 items with one of three possible labels as a gold label. Each item was given seven different annotations by a subset of 20 fictive annotators. A number of the 20 annotators was assigned a high competence score, while the rest was assigned a competence score equal to that of an annotator who labels randomly. We then produced MACE competence estimations and sorted the predictions from highest to lowest estimated competence. We then compared the estimations with the gold competence ranking using Kendall's Tau. This was repeated for varying ratios of competent and random workers, from 6 out of 20 to 20 out of 20, and under different definitions of a competent worker, varying from annotating 70% to 80% to 90% percent of items correctly. We repeated the process 5 times per condition.

The average correlation between the actual and predicted competence rankings was $r_\tau = 0.56$. These results show that MACE can be used to identify the least reliable annotators and to obtain a reliable majority vote even in quite unfavourable circumstances. We therefore used it to remove between 0 and 14 annotators with low MACE competence scores per country subset of the ETTC, and 3 annotators with low competence scores for the GTTC. For the ETTC, we additionally collected two extra judgments per tweet for the countries with lowest agreement (Russia and South-Africa). We then labeled tweets for stance with the majority vote, and removed tweets from the corpus that received the label *cannot read / does not apply (x)*.

4.2.2.4 Agreement

We measured agreement for the stance annotation task using Krippendorff's alpha (Krippendorff, 2018). The average agreement between ETTC annotations is 0.58 (Table 4.6). For the GTTC, the number of remaining annotators was 25, and they had an agreement of 0.62. These agreement scores indicate higher quality for the GTTC than ETTC. Both scores are higher than the alpha value of 0.57 obtained for thec coomparable task of 3-class sentence-level valence annotation in Antoine et al. (2014), where the authors compare Krippendorff's alpha and other agreement metrics on emotion and opinion annotation.

We performed a manual inspection of a sample of GTTC tweets to identify common reasons for disagreement. Reasons appear to include the need for more context to interpret a

Subcorpus	#Tweets	#Workers	Agreement
France	638	39	0.55
Indonesia	477	27	0.58
Russia	754	66	0.49
South Africa	698	82	0.51
Turkey	692	53	0.62
United States	743	43	0.64
Total	4002	204	0.58

Table 4.6 Number of tweets, number of workers and stance annotator agreement (Krippendorff's alpha) for the ETTC after removing those annotators that were judged to be least reliable by MACE.

tweet and the potential presence of irony. Table 4.12 shows examples of GTTC cases where annotators disagreed.

Tweet text	Translation	Annotations
UNGLAUBLICH!!! Wichtige Nachricht von @AfD-Chef **Dr. Jörg Meuthen**!	INCREDIBLE!!! Important message from @AfD-Chef **Dr. Jörg Meuthen**	-1, x, 0, 1, 0, 0, 1
Dass sich **Merkel** das noch alles antut. Ich hätte der CDU/CSU den ganzen Schmarren schon hingeschmissen.	[Can't believe] that Merkel is doing all this to herself. I would have let CDU/CSU deal with all this nonsense on their own long ago.	0, 0, -1, 1, 1, 1, 0

Table 4.7 Examples of tweets with disagreeing annotations. Possible labels were: negative (-1), neither positive nor negative (0), positive (1) and cannot read / does not apply (x). In the first example, it is difficult to assess the stance of the tweet without knowing the content of the important message. In the second example, the tweet can be taken as an expression of sympathy, or as an ironic statement.

4.2.3 Corpus Statistics

The label distributions for ETTC stance annotations and naming form formalities are presented in Table 4.8 and Table 4.9 respectively. GTTC label distributions are presented in Table 4.10 and Table 4.11. The total number of ETTC tweets after removing unreadable tweets is 4002. The final size of the GTTC corpus is 1904. The largest stance category in both data sets is the category *negative* (2197 tweets in the ETTC, 965 tweets in the GTTC). This is in line with research that shows that tweets with negative appraisals towards political

Country Subset	Positive	Neither	Negative	Total
France	155	208	275	638
Indonesia	173	187	117	477
Russia	120	126	508	754
South Africa	150	140	408	698
Turkey	165	159	430	692
United States	743	119	459	743
Total	866	939	2197	4002

Table 4.8 ETTC stance label distribution.

Country Subset	FN	LN	FNLN	TLN	TFNLN	Total
France	10	377	117	80	54	638
Indonesia	15	134	167	50	111	477
Russia	54	442	122	74	62	754
South Africa	6	405	109	106	72	698
Turkey	4	440	124	82	42	692
United States	59	363	141	94	86	743
Total	148	2161	780	486	427	4002

Table 4.9 ETTC naming form distribution.

parties are more common and spread faster than tweets with neutral or positive appraisals (Dang-Xuan et al., 2013).

4.3 Correlation between Stance and Naming Formality

We now examine the relation between the naming/titling and stance annotations in the collected TTC and GTTC tweets. As stated in Section 4.1, sociolinguistic work suggests

Stance	#Tweets
Negative	965
Neither	372
Positive	567
Total	1904

Table 4.10 GTTC stance label distribution.

Naming Form	#Tweets
LN	834
FNLN	307
TLN	371
TFNLN	392
Total	1904

Table 4.11 GTTC naming form distribution.

that naming conveys information about the status of an entity and/or solidarity with an entity and that both are signalled through variation in formality (Brown and Ford, 1961; Allerton, 1996; Dickey, 1997). The dual social function corresponds to the two possible relations which including a null-hypothesis of no relation between naming and stance correspond to the following three hypotheses:

H0 There is no relationship between variation in naming form and stance.

H1 Naming primarily downplays or emphasises the political figure's status. Therefore, formality of naming is positively correlated with stance.

H2 Naming primarily conveys the degree of solidarity with the political figure. Therefore, formality of naming is negatively correlated with stance.

Table 4.12 gives examples of ETTC tweets which can be interpreted to support either H1 or H2, or which suggest the existence of alternative, context-specific interpretations such as sarcasm.

Function	**Stance**	**Form**	**Tweet text**
status	pos	TFNLN	*Dear **President Joko Widodo**, Happy Birthday. God bless you @jokowi*
status	neg	FN	*That's the truth!!! Double-standard **#Donald** at it again*
solidarity	pos	LN	*Duterte & **Widodo** are truly public servants. Saving their countries fr the menace of society.*
solidarity	neg	TLN	***President Trump** probably won't like next week's newsstands*
sarcasm	neg	TLN	*Of course, I know, everything is sweetness & light in the wonderful democratic Paradise of **President Erdogan**!*

Table 4.12 Possible examples of status indication and solidarity indication through naming in ETTC tweets, and an example of an alternative function.

Because stance is an ordinal variable with the values negative (-1), neutral (0) and positive (1), we can compute average stance values per naming form category and measure the correlation between these values and another ordinal variable. To convert the naming form labels to an ordinal value, we rank them based on the following criteria:

1. Naming forms that include a title are more formal than naming forms that do not.
2. Longer naming forms are more formal than shorter naming forms.
3. Surnames are more formal than given names.

For the ETTC naming forms, this means we rank values as follows: FN<LN<FNLN<TLN<TFNLN. The ranking for the GTTC naming forms is: LN<FNLN<TLN<TFNLN. These rankings

allow us to assign tweets a **naming formality score** from 0 (LN) to 4 (TFNLN) for the ETTC and 0 (LN) to 3 (TFNLN) for the GTTC. Due to the oversampling of title-containing tweets for both corpora, the naming formality score cannot be used to estimate the typical naming formality of English-language or German-language tweets. It only serves to allow us to compute the correlation between naming formality and stance.

4.3.1 ETTC Findings

First, we contrast the average stance of ETTC tweets that contain the title *President* with tweets that do not. Table 4.13 shows that for all country subsets, the stance of tweets without a title in their naming form is lower than the stance of tweets whose naming forms include a title. A Kruskal-Wallis test reveals that these differences are all statistically significant with at p<0.01.

Secondly, we determine the correlation between naming formality and stance in the ETTC. Table 4.14 shows that the average stance of ETTC tweets increases with each increase in formality, and a Spearman's rank-order correlation test confirms that a statistically significant positive correlation exists between naming formality and stance ($r_s(4002) = .32, p = .001$).

These findings support the *status* hypothesis which states that because naming primarily indicates status, formality in naming/titling is positively related to stance.

Country Subset	**Without title**	**With title**
France	-0.24 (504)	-0.01 (134)
Indonesia	0.03 (316)	0.28 (161)
Russia	-0.62 (618)	-0.03 (136)
South Africa	-0.50 (520)	0.03 (178)
Turkey	-0.58 (568)	-0.02 (124)
United States	-0.55 (563)	0.09 (180)
Total	-0.45 (3089)	0.05 (913)

Table 4.13 Average stance and in parentheses the absolute number of tweets without or with a title in their naming form.

4.3.2 GTTC Findings

First, we contrast the average stance of GTTC tweets that contain the title *Dr.* with those that do not. A Kruskal-Wallis test shows that title-containing GTTC tweets have statistically significantly higher stance ($\mu = 0.09$, $n = 763$) than non-title-containing tweets ($\mu = -0.41$, $n = 1141$) ($\chi^2(4) = 145.46, p < 0.01$).

Country Subset	**FN**	**LN**	**FNLN**	**TLN**	**TFNLN**
France	0.00 (10)	-0.29 (377)	-0.08 (117)	-0.04 (80)	0.04 (54)
Indonesia	-0.60 (15)	-0.03 (134)	0.14 (167)	0.08 (50)	0.37 (111)
Russia	-0.56 (54)	-0.71 (442)	-0.31 (122)	-0.26 (74)	0.24 (62)
South Africa	-0.50 (6)	-0.53 (405)	-0.40 (109)	-0.08 (106)	0.18 (72)
Turkey	-0.75 (4)	-0.67 (440)	-0.23 (124)	0.06 (82)	-0.17 (42)
United States	-0.80 (59)	-0.53 (363)	-0.50 (141)	0.15 (94)	0.03 (86)
Total	-0.63 (148)	-0.52 (2161)	-0.21 (780)	-0.02 (486)	0.14 (427)

Table 4.14 Average stance and in parentheses the absolute number of tweets by country subset and naming form from least to most formal.

Secondly, we assess the relation between naming formality and stance in the GTTC. Table 4.15 shows that the average stance of GTTC tweets is lowest for tweets with the lowest naming form formality, and highest for tweets with the highest naming form formality. A Spearman's rank-order correlation test confirms that a statistically significant positive correlation exists between naming formality and stance in the GTTC ($r_s(1904) = .38, p < 0.01$). This is a somewhat higher correlation than that found for the ETTC ($r_s(4002) = .32$).

We observe a moderate, statistically significant positive correlation between formality and stance The stance of title-containing and non-title-containing GTTC tweets and the correlation between stance and naming formality in the GTTC both suggest that the status-indicating function of naming and titling is stronger than the solidarity-indicating function. This means support for H2 comes not only from English-language tweets containing the professional title *President*, but also from German-language tweets containing the academic title *Dr.*.

4.3.2.1 Observations on additional titles and title combinations

As explained in Section 4.2.1.2, GTTC naming form labels are the result of a labeling strategy that categorises naming forms into four general classes. Underneath this four-way categorisation, the GTTC contains a total of 19 different combinations of names and titles, most of them represented by only a handful of tweets.

The titles that appear other than Dr. are *Frau* (Ms.), *Herr* (Mr.), the professional titles *Bundeskanzlerin* and *Kanzlerin* (Chancellor), used only for Angela Merkel (CDU/CSU), and the academic title *Professor*, used only for Jörg Meuthen (AfD). Table 4.16 shows that the appearance of the generic titles *Frau* and *Herr* in tweets is associated with negative stance. Table 4.17 shows that the appearance of professional and academic titles, here including *Dr.* for comparison, is associated with positive stance.

Politician Subset (Party Subtotals)	**LN**	**FNLN**	**TLN**	**TFNLN**
Sahra Wagenknecht	0.0 (17)	-0.12 (8)	1.0 (1)	0.86 (7)
Gregor Gysi	0.5 (2)	0.75 (4)	1.0 (1)	-0.25 (4)
Dietmar Bartsch	-0.5 (2)	1.0 (1)	-1.0 (1)	0.5 (2)
Die Linke	0.0 (21)	0.23 (13)	0.33 (3)	0.46 (13)
Robert Habeck	-0.23 (13)	0.0 (7)	-1.0 (1)	0.5 (2)
Anton Hofreiter	-1.0 (3)	-0.33 (3)	-1.0 (1)	0.0 (3)
Simone Peter	-1.0 (1)	0.0 (1)	N/A (0)	0.0 (1)
Bündnis 90/Die Grünen	-0.41 (17)	-0.09 (11)	-1.0 (2)	0.17 (6)
Franziska Giffey	-0.38 (21)	0.3 (30)	N/A (0)	0.81 (52)
Karl Lauterbach	-0.73 (15)	-0.75 (4)	-0.5 (8)	0.5 (2)
Frank-Walter Steinmeier	-0.45 (11)	0.57 (7)	N/A (0)	1.0 (2)
Karamba Diaby	1.0 (1)	0.43 (7)	N/A (0)	0.71 (7)
Katarina Barley	-0.5 (2)	0.5 (2)	N/A (0)	0.0 (3)
SPD	-0.48 (50)	0.28 (50)	-0.5 (8)	0.76 (66)
Angela Merkel	-0.66 (385)	-0.19 (115)	-0.39 (70)	-0.28 (79)
Wolfgang Schäuble	-0.18 (11)	0.25 (4)	0.5 (4)	0.54 (13)
Gerd Müller	0.0 (2)	0.33 (3)	0.33 (3)	0.64 (14)
Norbert Lammert	1.0 (2)	1.0 (5)	N/A (0)	0.5 (6)
Ursula von der Leyen	0.0 (4)	0.0 (2)	-1.0 (1)	1.0 (1)
Helge Braun	0.0 (1)	0.0 (1)	N/A (0)	0.25 (4)
Peter Tauber	-1.0 (2)	-1.0 (1)	-1.0 (1)	0.0 (2)
Thomas de Maizière	N/A (0)	0.0 (2)	N/A (0)	0.0 (2)
Kristina Schröder	0.0 (1)	N/A (0)	-1.0 (2)	N/A (0)
CDU/CSU	-0.63 (408)	-0.12 (133)	-0.35 (81)	-0.01 (121)
Alice Weidel	-0.49 (213)	-0.11 (70)	-0.36 (203)	0.56 (100)
Alexander Gauland	-0.49 (47)	-0.07 (14)	0.61 (41)	1.0 (1)
Jörg Meuthen	-0.54 (76)	0.22 (9)	0.0 (32)	0.43 (76)
Frauke Petry	0.0 (2)	-0.29 (7)	-1.0 (1)	0.44 (9)
Petry	0.0 (2)	-0.29 (7)	-1.0 (1)	0.44 (9)
AfD/Petry	-0.5 (338)	-0.09 (100)	-0.18 (277)	0.51 (186)
Total	-0.55 (834)	-0.03 (307)	-0.22 (371)	0.38 (392)

Table 4.15 Average stance and in parentheses the absolute number of tweets by politician (with party subtotals) and naming form from least to most formal.

The most frequent combination of titling forms is *Frau Dr.* This combination occurs no less than 295 times. In the far majority of cases, the combination is used in reference to controversial far-right politician Alice Weidel (AfD) (71%)). Table 4.18 shows examples of the*Frau Dr.*-combination in context. Tweets containing this combination have low stance

(-0.35, n=295), but the stance is not significantly different from that of female-target tweets without *Frau Dr.* (-0.28, n=1116).

	With Form		Without Form	
Title	**#Tweets**	**Stance**	**#Tweets**	**Stance**
***Frau**	352	-0.40	1059	-0.25
Herr	25	-0.36	468	0.04

Table 4.16 Number and average stance of GTTC tweets that do or do not contain *Frau* (tweets with a female target) and *Herr* (tweets with a male target). In rows marker with "*", the difference is statistically significant ($p < 0.01$).

	With Title		Without Title	
Title	**#Tweets**	**Stance**	**#Tweets**	**Stance**
***Dr.**	763	0.09	1141	-0.41
***Professor**	94	0.31	99	-0.35
***B/K**	70	0.13	579	-0.58

Table 4.17 Number and average stance of GTTC tweets that do or do not contain *Dr.* (all tweets), *Bundeskanzlerin* or *Kanzlerin* (B/K) (tweets with target *Angela Merkel*) and *Professor* (P) (tweets with target *Jörg Meuthen*).

4.4 Impact of Political Orientation and Target Gender

4.4.1 Political Orientation

Sociolinguists have pointed out that while there are standard patterns for the usage of names and titles, the exact meaning of a particular choice of name or title also depends on context factors (Dickey, 1997). For German specifically, sociolinguists have noted that use of the second person pronoun, which is a form of address just like names and titles are, is affected by the political orientation of the speaker (Besch, 1998; Hickey, 2003; Clyne et al., 2006). Speakers that are left-leaning use the informal pronoun *Du* more frequently to signal equality and solidarity, whereas speakers that are right-leaning make more frequent use of the formal pronoun *Sie* out of respect for the social hierarchy. This observation echoes observations of moral psychologists which state that there is a right-left divide in whether hierarchies are viewed as an asset to conserve or an injustice to resist (Graham et al., 2009).

Tweet Text & Translation	**Stance**
Frau Dr. Alice Weidel machen sie weiter so. Sie sprechen Menschen mit Sachverstand und dem Herz am rechten Fleck an. *Dr. Alice Weidel keep it up. You approach people with expertise and the heart in the right place.*	Positive
Frau Dr. Alice Weidel spricht endlich mal die Wahrheit aus, die die Regierungsparteien und Gutmenschen vehement leugnen. *Dr. Alice Weidel finally speaks out the truth which the government parties and do-gooders vehemently deny.*	Positive
Frau Dr. Wagenknecht war heute im Bundestag in Bestform! Teilen, teilen, teilen. *Dr. Wagenknecht was in top form in the Bundestag today! Share, share, share.*	Positive
Sie sind widerlich, **Frau Dr. Weidel**. *You are repulsive, Dr. Weidel*	Negative
Mit Verlaub **Frau Dr. Weidel** - Sie haben gewiss nicht alle Latten am Zaun. *With all due respect Dr. Weidel - you've clearly got a few screws loose.*	Negative
Alles Gute zum Geburtstag, **Frau Dr. Angela Merkel**! Und viel Spaß im Ruhestand! #frischesbayern #fdp #liberal *Happy birthday, Dr. Angela Merkel! And have fun in retirement! #freshbavaria #fdp #liberal*	Negative

Table 4.18 Examples of tweets with the title combination *Frau Dr.* and either positive or negative stance.

In this section, we examine whether political orientation influences the status-indicating function of naming that we have observed thus far. Because of the existing literature suggesting an impact of political orientation on pronoun use in the German language, we perform our tests on GTTC tweets. We hypothesize that the strength of the positive relation between formality and stance found for GTTC tweets overall is weaker among tweets from left-leaning Twitter users than among tweets from right-leaning Twitter users.

We conduct two studies which differ in how we determine the political leaning of the source of the tweet. For the first study, we use a proxy for political leaning that is less reliable, but can be applied to all 1904 tweets. For the second study, we use human annotations for political leaning that are more reliable, but more costly, and therefore only applied to an addendum to the GTTC that consists of a smaller number of tweets, the construction of which is discussed in the same section (Section 4.4.1.2).

4.4.1.1 Proxy Study

For the proxy study of all 1904 GTTC tweets, we infer the most likely political leaning of a tweet's source from its stance label and stance target. We then separate the corpus into a left-leaning and right-leaning segment, and compare the strength of the association between naming formality and stance in the two segments.

To infer the most likely political leaning of tweet writers, we take three steps. First, we discard tweets with neutral stance, and retain tweets that praise or criticize their stance target. In this step, we also remove tweets whose stance target is a politician from a government party (CDU/CSU, SPD), because these targets likely attract criticism from both sides of the political spectrum. This leaves 826 tweets with negative or positive stance towards politicians from opposition parties.

As a second step, we divide the selected tweets into tweets whose target is from a right-leaning opposition party and tweets whose target is from a left-leaning opposition party. Based on the political spectrum positioning given in Table 4.4, this corresponds to tweets with targets who are either a current or former member of the right-leaning AfD, or a member of the left-leaning B90/Grün and Die Linke.

Lastly, we make an inference based on the two-fold assumption that:

- tweets with negative stance towards right-leaning politicians and positive stance towards left-leaning politicians are from users who are **likely left-leaning (LLL)**, and
- tweets with negative stance towards left-leaning politicians and positive stance towards right-leaning politicians are from users who are **likely right-leaning (LRL)**.

This gives us two tweet categories: the category LLL which contains the stance directions anti-right (**AR**) and pro-left (**PL**), and the category LRL which contains the stance directions anti-left (**AL**) and pro-right (**PR**) (Table 4.19).

User	**Size by Stance**		
Leaning	**Negative**	**Positive**	**Total**
LLL	449 (AR)	31 (PL)	480
LRL	30 (AL)	316 (PR)	346
Total	479	347	826

Table 4.19 Number of GTTC tweets with likely left-leaning (LLL) or likely right-leaning (LRL) users and pro-left (PL), anti-right (AR), pro-right (PR) or anti-left (AL) stance.

If our proxy for identifying political orientation distinguishes between left-leaning and right-leaning discourse well enough, and if indeed names and titles are less status-indicating in left-leaning discourse, the LLL and LRL group should differ in the strength of their associations between naming formality and stance. Because we removed neutral tweets, stance here is a binary variable, and we can not compute a formality-stance correlation to compare the LLL and LRL groups like we did in Section 4.3.1 and Section 4.3.2. Instead, we will compare the discrepancy in formality between positive and negative tweets within each of these groups.

The difference in formality, which we will call the **formality gap**, can be measured using the naming formality score defined in Section 4.3 on page 62. A larger gap indicates a stronger status-indicating function and a smaller gap indicates a weaker status-indicating function. As a point of reference: the average formality score of the original GTTC (including neutral tweets) is 0.75 for negative tweets and 1.77 for positive tweets, meaning the absolute formality gap in the GTTC overall is $\delta = 1.02$, and the relative formality gap is 111% of the value for negative tweets.

Table 4.20 gives the formality gaps for the LLL and LRL tweets. We observe that the formality gap between positive and negative LLL tweets is much smaller ($\delta = 0.34, 37\%$) than the gap between positive and negative LRL tweets ($\delta = 1.24, 207\%$). This confirms our hypothesis that naming forms in tweets from left-leaning users are less strongly status-indicating than naming forms in tweets from right-leaning users.

User	**Formality by Stance**		**Formality**	
Leaning	**Negative**	**Positive**	**Gap (δ)**	**Total**
LLL	0.92 (AR)	1.26 (PL)	0.34 (37%)	0.94
LRL	0.60 (AL)	1.86 (PR)	1.24 (207%)	1.75
Total	0.90	1.8	0.90 (100%)	1.28

Table 4.20 Average naming formality in GTTC tweets with likely left-leaning (LLL) or likely right-leaning (LRL) users and pro-left (PL), anti-right (AR), pro-right (PR) or anti-left (AL) stance, and the naming formality gap as an absolute number and a percentage of negative tweets.

4.4.1.2 Addendum Study

There are several drawbacks to the proxy study of Section 4.4.1.1. Firstly, the assumption is made that users expressing criticism of a politician are criticising their opposites on the political spectrum. This assumption is likely too strong. It has been shown on newspaper data that bias towards one side of the political spectrum manifests in a tendency to cover entities from that side of the spectrum less negatively, but not exclusively positively (Fan et al., 2019). Manual inspection of the GTTC Addendum shows that this may also be the case in tweets about politicians. We found examples of criticism of the centre-left from the more radical left, and criticism of former AfD member Frauke Petry from current AfD-supporters. We also found many tweets which praise Sahra Wagenknecht but profess in the same tweet to dislike her party, Die Linke.

Secondly, GTTC tweets were sampled so as to over-represent title-containing tweets (see Section 4.2.1), which may have a confounding effect on the formality gaps we observed. Finally, the GTTC contains considerably more tweets with right-wing politician targets (AR and PR stance subgroups) than left-wing politician targets (AL and PL subgroups) and the AL and PL subgroups in the study are consequently smaller than ideal for empirical research.

To address these drawbacks, we conduct a second study on an additional collection of data. This addendum has a natural distribution of naming forms and a more balanced distribution of left-wing and right-wing targets, and is annotated with manual annotations of user orientation.

To construct the addendum, we collected tweets mentioning each of the four right-wing politicians (AfD members Alexander Gauland, Jörg Meuthen and Alice Weidel, and independent politician Frauke Petry) and each of the six left-leaning politicians (members of B90/Grün or Die Linke: Sahra Wagenknecht, Gregor Gysi, Dietmar Bartsch, Robert Habeck, Anton Hofreiter, and Simone Peter). For each target entity, we collected the three most popular non-news tweets of each month of 2018, which were then stance-annotated by an outside native speaker annotator using the same prompts as described in 4.2.2.2. Tweets which had no clear stance direction were discarded. Tweets were grouped by their user to provide annotators with as much relevant context as possible, and each of these 216 user groups was shown to two expert, native speaker annotators. The two annotators saw users' tweets along with the user name, user location and user description. Annotators were encouraged to follow a link to the user profile for more context if necessary. They were asked to decide for each user whether they would categorize them as *left-leaning* (**LL**) or as *right-leaning* (**RL**). Agreement between the two annotators was 0.85 (Cohen's kappa (McHugh, 2012)). We kept only tweets by users whose political orientation was unambiguous, i.e. given the same annotation by both annotators.

The resulting GTTC Addendum consists of 296 tweets with an average stance of 0.18 (compared to -0.16 in the GTTC) and an average naming formality of 0.60 (compared to 1.28 in the GTTC, where titles were oversampled). Table 4.21 shows that the distribution of the four stance direction groups is more balanced in the addendum compared to the distribution shown in Tabel 4.19.

In the GTTC Addendum, the formality gap between the stance direction groups of LL tweets is much smaller ($\delta = 0.14, 28\%$) than the formality gap between the stance direction

User	Size by Stance		
Leaning	Negative	Positive	Total
LL	116	66	182
RL	59	55	114
Total	175	121	296

Table 4.21 GTTC Addendum tweets by user leaning (left-leaning (LL) or right-leaning (RL)) and stance.

groups of RL tweets ($\delta = 0.65, 176\%$) (Table 4.22). Here, too, the data suggests naming formality is more weakly associated with stance in left-leaning discourse than in right-leaning discourse. This is in line with our expectations, which were based on Moral Foundations Theory (Graham et al., 2009) and state that right-leaning users likely place more value on respect for status and have more positive associations with indicators of higher social status such as the title *Dr.*.

User	Formality by Stance		Formality	
Leaning	Negative	Positive	Gap (δ)	Total
LL	0.50	0.64	0.14 (28%)	0.55
RL	0.37	1.02	0.65 (176%)	0.68
Total	0.46	0.81	0.35 (76%)	0.60

Table 4.22 Average naming formality in the GTTC Addendum across left-leaning (LL) or right-leaning (RL) users and negative or positive stance , and the naming formality gap as an absolute number and a percentage of negative tweets.

4.4.2 Target Gender

Prevous work has suggested that titles are used more frequently for male than for female targets (Takiff et al., 2001; Uscinski and Goren, 2011). This work was conducted in an English setting, and we are interested whether the same finding can be replicated for German-language mentions of political entities with doctoral titles. Verifying the difference in titling frequency in reference to male and female politicians does not require stance labels, which allows us use a large amount of unlabeled data, in the form of the the Social Media Monitoring Corpus[6]. This corpus tracks mentions of German politicians on Twitter from July 2017 onwards.

[6] mediamonitoring.gesis.org

To assess titling frequency for female and male politicians in the SMMC, we take the subset of GTTC politicians that appears in the SMMC and request all tweets which refer to them. We use GTTC politicians because we know them to be prominent enough to attract some social media attention and we know them to possess doctoral titles. The subset of politicians we submitted queries for consisted of 13 male (**M**) politicians and 8 female (**F**) politicians: Alexander Gauland (M), Alice Weidel (F), Angela Merkel (F), Anton Hofreiter (M), Dietmar Bartsch (M), Frauke Petry (F), Gregor Gysi (M), Helge Braun (F), Karamba Diaby (M), Karl Lauterbach (M), Katarina Barley (F), Peter Tauber (M), Sahra Wagenknecht (F), Ursula von der Leyen (F) and Wolfgang Schäuble (M).

We capped mentions to a maximum of $10,000$ to reduce skew towards the most high-profile politicians, retaining $772,909$ mentions, of which 55% of female targets. Table 4.23 shows the distribution of title-containing and non-title-containing tweets by target gender, and shows that the percentage of title-containing tweets is higher for female targets. A chi-square significance tests shows that the observed distribution is not significantly different from the distribution that is expected if target gender has no impact on title usage.

Gender	**Without title**	**With title (%)**	**Total**
F	415362	6415 (2%)	421777
M	347449	3683 (1%)	351132
Total	762811	10098	772909

Table 4.23 Quantities of tweets with and without title for mentions of female and male targets.

4.5 Discussion

The studies presented in this chapter are suggestive of a possible framing effect of naming/titling that is modulated by political orientation, but have several limitations that are worth noting.

The automatic string matching strategy used to label naming forms in the ETTC and GTTC does not distinguish between address and reference. This means it applies the same labels regardless of whether the naming form is used as a term of address (e.g. “Making things "Great Again" huh #Donald?”) or as a reference form (e.g. “#Donald just cant handle competing for the title.”). Studying these types separately would require additional manual annotation, because using the `@`sign, the symbol which directs a tweet to an account, as a heuristic for labeling would return both false positives (*@jokowi we suport Mr Joko Widodo*) and false negatives (*Happy Birthday President Mr. Joko Widodo*). In addition, the distinction between address and reference is not as clear for Twitter data as for face-to-face conversations, as many tweets mix both functions. Future work is needed to investigate whether naming forms in address vs. reference impact stance differently.

Additionally, both corpora consist of tweets from a relatively limited time span. This means the content of the tweets and therefore the naming used in them may be influenced by the occurrence of specific events (e.g. President Joko Widodo's birthday) and not generalize well. The ETTC has an additional weakness in that it contains only English-language tweets, but its entity targets are mostly from countries where English is not the primary language. English-language tweets about these presidents may be unrepresentative of local ways of referring to the president. They may also be more neutral in tone and may use (T)FNLN to be informative rather than respectful. These two factors along with certain other ones (the character limit of tweets, the often unspecified audience) increase the chance that the primary function of naming is lost among noise. It is therefore interesting to nevertheless observe a statistically significant trend across within the ETTC's country subsets and across the two data sets that informal naming of presidents co-occurs with perceived hostility, while formal naming co-occurs with perceived supportiveness of a tweet.

4.6 Summary and Future Work

In the chapter, we presented an analysis of naming/titling of political entities in social media, its relation to stance, and its interaction with the context factors source political orientation and target gender. Our analysis reveals a positive relation between naming/titling formality and stance in two corpora with distinct languages and naming form sets. These findings confirm sociolinguistic claims that naming marks status and expresses respect that has not previously been shown in a large, quantitative study, nor for social media texts. They also constitute the first quantitative support for a possible framing effect of naming variation.

In addition to a general status-indicating function of naming, we also report findings on the role of the political orientation of tweets' source and the gender of tweets' target entities. We find that the positive association between formality and stance is weaker in left-leaning discourse than in right-leaning discourse on two differently sampled German corpora, one of them directly annotated for source political orientation. This is the first quantitative evidence of an impact of the writer's political orientation on naming/titling, and of an interaction between entity framing and the differences in value systems suggested by moral foundations theory. We do not find evidence of gender bias in naming/titling: our comparison of the use of an academic title for politicians in a large collection of German tweets showed no difference between the frequency of titling for male and female politicians.

Future work can extend this research by creating and analysing data collections that cover a longer time span than the ETTC and GTTC, and with annotations that allow experiments on the impact of factors other than source political orientation and target gender. These include the age and gender of tweets' sources, and whether the politician in question is being referred to or whether they are being addressed.

Future work can also investigate the degree to which formal naming is used sarcastically. Our data shows that despite highly formal names occurring more frequently in positive tweets, they certainly also occur in tweets with negative stance. Since the use of honorifics can be indicative of sarcasm (Liu et al., 2014), it is worth investigating whether the use of titles alongside explicit negative stance should be interpreted as sarcasm. Also interesting is whether sarcastic use plays a role in causing the weaker positive association with formal naming in left-leaning discourse.

Finally, NLP work in the field of Stance Detection can use our corpora as further training and testing material for experiments on the detection of stance towards politicians on social media. Experiments in Mohammad et al. (2016) show that cross-target stance detection is more challenging than in-target stance detection. Future work may wish to examine whether Stance Detection systems trained for a subset of GTTC or ETTC entities can also predict stance for held-out entities.

Chapter 5

Detection of Informational Entity Framing

5.1 Introduction

So far we have concentrated on how word-level variation influences perceived opinions towards entities. Doing so contributes to a theoretical understanding of how mentioning objectively verifiable attributes may impact perceived stance. For real-world applications, however, it is necessary to look beyond word-level variation and examine Entity Framing on the level of larger spans of text.

To help journalists ensure and to help readers assess the neutrality of news texts, we need systems that can highlight spans of non-neutral reporting on entities. Part of this task can be solved by Bias Detection systems which identify reference forms and modifiers that are associated with a certain world-view. Another part can be solved by Stance Detection systems which find explicit opinions towards entities as well as terms with a negative or positive connotation. But what neither of these systems recognise is neutrally formulated text that, while factual, guides attitudes towards entities in a positive or negative direction.

Framing through neutrally formulated text is prevalent in news, because the purpose of news texts is not only to report on newsworthy events, but also to place those events in a broader context (Fan et al., 2019). Intentionally or not, the inclusion of background information that contextualises events influences not only the perception of events, but also of the entities involved in them. For example, journalists may include allusions to the motivations or goals behind actions taken by entities, or they may give a voice to third-party interpretations of an entities' behaviour. Writers may compare entities' current behaviour with previous behaviour, or they may describe attributes of the entities that were judged to be

Idx	Sentence	Inf	Src	ID
1.1	Former Arkansas Gov. Mike Huckabee announced Tuesday he is running for president [. . .].	0	FOX	53fox00
1.2	Mr. Huckabee opposes same-sex marriage, suggesting as recently as February that homosexuality is a lifestyle choice akin to drinking or swearing.	1	NYT	53nyt10
1.3	He was the longest-serving Arkansas governor, from 1996 to 2007.	1	HPO	53hpo15

Table 5.1 Example instances from a story in the BASIL corpus of informational bias (Fan et al., 2019) with their informational bias label (inf), news source (src) and BASIL ID.

relevant to the main event in more detail than others that were not. Fan et al. (2019) call this Informational Bias and define it as “sentences or clauses that convey information tangential, speculative, or as background to the main event in order to sway readers’ opinions towards entities in the news”. We follow this definition, but refer to the phenomenon as Informational Entity Framing (IEF).

For an example of IEF, see Instances 1.2 and 1.3 in Table 5.1. Consider especially Instance 1.3. This instance contains no subjective language. Seen on its own, it is simply stating a fact. However, human annotators of the corpus from which it was taken (Fan et al. (2019)’s BASIL corpus) judged that it is an instance of IEF because the mentioned fact reflects positively on the target entity Mike Huckabee in the context of an announcement that he is running for president. Note that one can also imagine contexts where this fact reflects negatively on Mike Huckabee. An example would be a discussion of a disconnect between older Republican candidates and a new generation of more progressive voters.

There is very little work on the automatic detection of IEF and the work that does exist has only attempted to recognise it in isolated sentences. This is problematic, because while some instances of IEF are recognisable even outside of their context (e.g. quotations from third parties that contain explicit opinions), others, like Instance 1.3 from Table 5.1, are mere statements of facts that do not raise suspicions of IEF outside of their context. Fan et al. (2019) emphasize that, unlike more commonly studied kinds of bias, IEF label assignments depend very heavily on context. For this reason, the human experts who gave IEF annotations for sentences saw them in the context of not only the article they appeared in but also of several news articles on the same event.

In this chapter, we explore whether and how the integration of context in IEF detection boosts performance. We hypothesise that, as evidenced by BASIL annotators’ instructions to read entire articles and even triples of articles before identifying IEF spans, sentences should not be labeled for IEF in isolation, but should be contextualized. To test this hypothesis, we compare an array of context-inclusive approaches. We first establish a sentence-only baseline that bases its predictions only on representations of input sentences. We then define three kinds of context that may be relevant for IEF and experiment with one or more promising techniques for integrating each kind of context into IEF detection. Because we are testing approaches that have never been applied to IEF before, we limit our scope to sentence-level binary IEF classification systems. Performing more fine-grained labeling at the level of the token implies both a heavier computational load and increased complexity. We delegate this more complex variant of the task to future work. All code for all experiments in this chapter has been made publically available[1].

5.2 The BASIL Corpus

All our context-integration experiments are conducted on the first and only data set with Informational Entity Framing (IEF) annotations: the BASIL (Bias Annotation Spans on the Informational Level) dataset (Fan et al., 2019). This corpus contains 100 triples of news articles from Fox News (FOX), the New York Times (NYT) and the Huffington Post (HPO). Each triplet covers a single news event, and is labeled with event and entity annotations. Documents are labeled additionally with stance annotations, and sentences are labeled with IEF annotations, as well as lexical bias annotations that mark text that displays bias through word choice. The IEF and bias annotations come with span ranges and span-level target, polarity and aim annotations. Because the IEF and bias annotations have span starts and ends, the corpus can be used for both token-level classification and for sentence classification. For binary sentence-level classification, a sentence is labeled as containing IEF if it contains at least one span of IEF.

The BASIL corpus' comprises 10 triplets of articles from each year of the 2010's that were selected through a combination of automatic alignment and manual inspection. After selecting triplets, Fan et al. (2019)'s authors added lists of the main entities and main events covered in each triplet as triplet-level annotations. The authors then asked human annotators to read triplets as a unit and provide document-level and sentence-level annotations. On the document level, the annotators determined articles' stance towards the main event and the main entities of the triplet. They also ranked the triplets' articles from most left-leaning to most right-leaning. On the level of the sentence, the annotators identified spans of text that contained lexical bias and/or IEF. They labeled each identified span with a target, a polarity

[1] https://github.com/vdenberg/context-in-informational-bias-detection/blob/master/experiments/finetune_plm.py

Polarity	N
Negative	990
Positive	259
All	1221

Table 5.2 Polarity labels of BASIL IEF instances.

Aim	N
Direct	1095
Indirect	126
All	1221

Table 5.3 Aim labels of BASIL IEF instances.

Quote	N
No	641
Yes	608
All	1221

Table 5.4 Quote labels of BASIL IEF instances.

label, and an aim label. The aim of a span of bias or framing refers to whether the polarity label needs to be inferred from the text (indirect aim) or not (direct aim). They also stated for each span whether it is part of a quotation or not. For an example of a BASIL entry and an IEF span, see Appendix B.

After annotation, Fan et al. (2019)'s authors determined the corpus' gold standard by organising discussions between the annotators whose aim was to resolve disagreements and remove personal biases from the annotation. They assessed Inter-annotator Agreement using the F1-score of span overlaps between pairs of annotators. The agreement scores are on the lower end (F1 $= 0.34$ for IEF, F1 $= 0.14$ for lexical bias), but the average agreement between individual annotations and the aggregated gold standard is on the higher end, especially for IEF (F1 $= 0.70$ for IEF, F1 $= 0.56$ for lexical bias). The corpus contains 7,977 sentences, seven less than reported in the original paper because we encountered seven instances during pre-processing with a length of zero. Tables 5.2, 5.3 and 5.4 give the distribution of polarity, aim and quote labels of IEF instances. The total number of IEF-containing sentences is 1221. Most IEF instances (990) have negative polarity, and the vast majority of IEF instances (1095) has a direct aim. Almost half of IEF instances (608) appear in a quote.

5.3 Sentence-only Baseline

The state-of-the-art in IEF detection at the time of writing is the sentence-only BERT-based approach by Fan et al. (2019). Fan et al. (2019) train BERT on a training and test split that they made available upon request. We noticed that this split randomly divides individual sentences across a training, development and test set. Splitting data this way isolates target sentences from other sentences in the same article, as well as from other articles covering the same event.

We do not want to isolate sentences in data splits for our experiments because it is a) contrary to our goal of considering sentences within their context and can b) be considered a type of information leakage from train to test. Knowing of some sentences in an article that they are biased might help models recognise similar sentences from the same article or another article on the same topic. Moreover, this scenario is an unrealistic setting for potential use of an IEF detection system. A hypothetical user of such a system is likely to want to judge for new, unread articles whether they contain IEF, not for a subset of sentences from archived articles.

Because of the drawbacks of the split used in Fan et al. (2019), henceforth the **Sentence Split**, we propose the **Story Split**, a split where data is distributed across training and test sets in a way that ensures triples of article (stories) appear either during training or during testing but not both. We report fine-tuning results on the Sentence split for the purpose of reproducing and comparing to Fan et al. (2019) and for the purpose of observing the difference with results on the Story Split. The Sentence Split consists of 6,819 training instances, 758 development instances and 400 test instances. The **Story Split** is a 10-fold cross-validation setting where stories (triplets of articles) never appear in both a train and non-train section. Sizes of folds in the Story Split vary slightly because of variation in the length of articles. Each consists of around 6,400 sentences designated for training, 780 for development and 790 for testing.

Because of improved scores observed by RoBERTa compared to BERT on other tasks (Liu et al., 2019b), we report fine-tuning results of RoBERTa as well as BERT. We train both models 5 times per setting, each time with a different random seed. We report precision, recall and the F1-score (with standard deviation across seeds) for the IEF-containing class and test significance of differences in performance with an independent t-test ($p < 0.05$). Further training details included hyperparameters selected with hyperparameter tuning are provided in Appendix C.

We provide results for token-level classification to show the effect of changing from a Sentence Split to a Story Split on this task, but do not provide context-integration results for token-level classification in later sections. The reason for this, as stated in Section 5.1, is that we consider it premature to add this layer of computational complexity before more is known about whether and how context-integration benefits IEF detection.

5.3.0.1 Baseline results

In line with the prediction that the Sentence Split introduces leakage of information, classification performance is several F1-score points higher on the Sentence Split than on the Story Split in all settings (Table 5.5). The difference in F1-scores is largest for sentence classification with RoBERTa (49.89 vs 42.16).

We observe large improvements in performance of RoBERTa over BERT in all settings. Best performance on sentence classification in Fan et al. (2019) was 43.27 by BERT. In our set-up, BERT's F1-score stands at 38.26 on the Sentence Split and RoBERTa's at 49.89. On

Task	Set-Up	Model	Precision	Recall	F1-score
Token	Sentence Split	Fan et al. (2019)	25.56	14.78	18.71
		BERT	12.42 ± 1.31	28.31 ± 3.18	17.23 ± 1.61
		RoBERTa	36.10 ± 4.51	32.41 ± 2.92	34.03 ± 2.81
	Story Split	BERT	12.85 ± 1.06	22.12 ± 1.60	14.60 ± 0.91
		RoBERTa	32.44 ± 2.04	27.73 ± 1.54	29.86 ± 1.25
Sentence	Sentence Split	Fan et al. (2019)	43.87	42.91	43.27
		BERT	46.44 ± 2.51	33.0 ± 7.21	38.26 ± 5.29
		RoBERTa	47.55 ± 2.92	52.67 ± 6.41	49.89 ± 4.06
	Story Split	BERT	40.54 ± 0.62	31.43 ± 1.56	35.39 ± 1.02
		RoBERTa	43.12 ± 1.03	41.29 ± 1.37	42.16 ± 0.30

Table 5.5 Baseline performance without context for token and sentence classification. Sentences are divided into training and non-training sets by sentence or by story. We report standard deviations across 5 seeds for our models. Fan et al. (2019) report a minimum standard deviation of 3.36 and maximum of 12.44 for theirs.

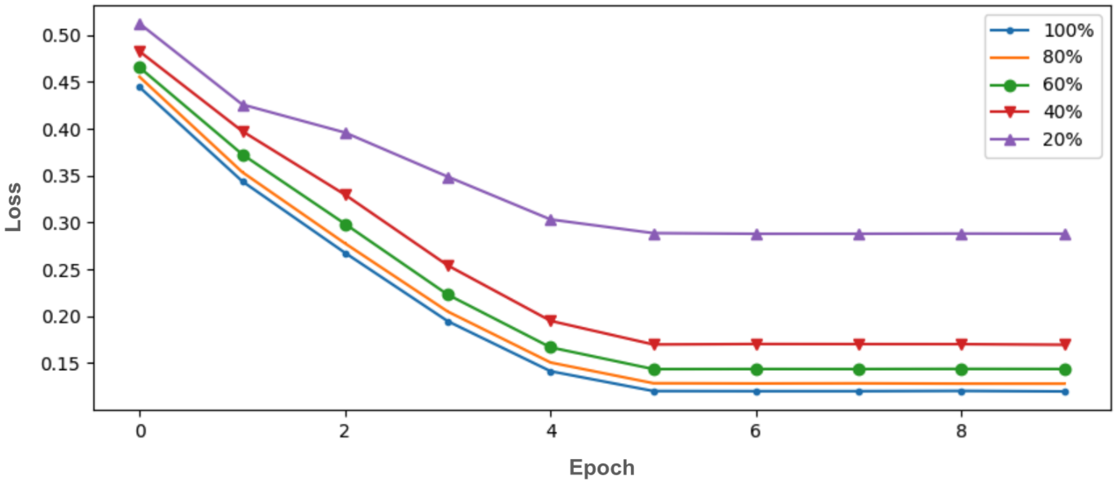

Fig. 5.1 Sentence-only baseline loss values on training data per epoch for different amounts of training data.

the Story Split the difference is also large: from 35.39 by BERT to 42.16 by RoBERTa. Like Fan et al. (2019), we observe lower performance for token classification than for sentence classification (29.86 vs 42.16 (RoBERTa on the Story Split)).

A final observation is that the standard deviations of F1-scores are higher for Sentence Split settings than Story Split settings. For example, the standard deviation of the F1-scores for sentence-level classification is 5.29 for BERT and 4.06 for RoBERTa for the Sentence

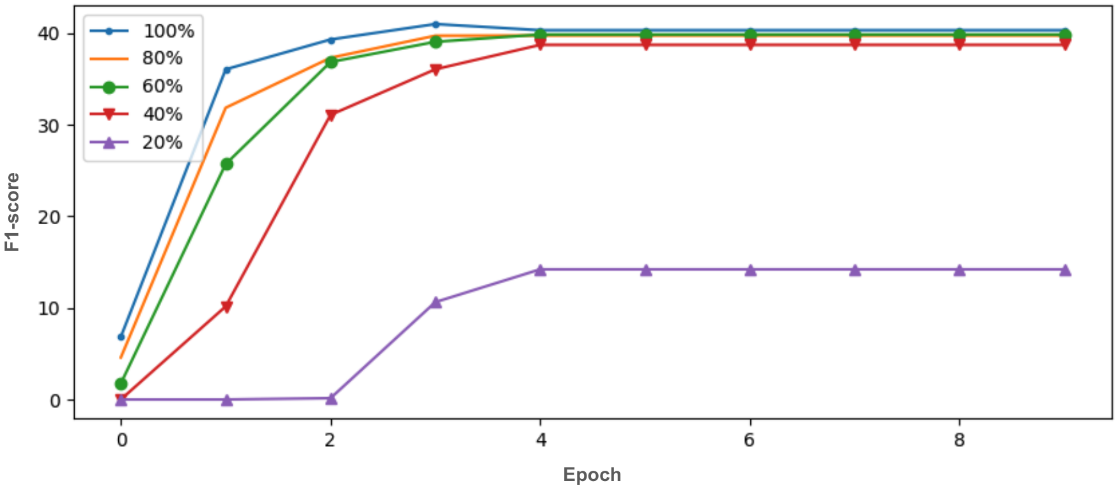

Fig. 5.2 Sentence-only baseline F1-scores on validation data per epoch for different amounts of training data.

Split versus 1.02 and 0.30 for the Story Split. This suggests that although systems detect some additional useful patterns in the Sentence Split, they learn them less reliably. It is known that PLMs can be sensitive to changes in seeds (Mosbach et al., 2020), so the increased stability of the Story Split results is an important benefit of splitting data this way.

The baseline setting that we select as a point of comparison for new approaches is the sentence-level RoBERTa system trained on the Story Split. Figures 5.1 and 5.2 show how training loss values and validation F1-scores of this setting respond to increasing the size of the training data. Figure 5.2 shows this setting converges after roughly 4 epochs. The highest validation F1-score when training on 100% of the data is 3% better than the best score when training on 80% of the data, which suggests it would be further improved if the data set were larger.

5.4 IEF Detection with context beyond the sentence

We organize our exploration of context-integration around a breakdown of the broad notion of “context” into different context types. We define three kinds of context that are relevant for IEF, increasing in scale and in distance to the target sentence: article context, event context, and domain context.

1. **Article context**. IEF in news by definition appears within a news article consisting of other pieces of text. The text that immediately surrounds an input sentence may be helpful for disambiguating sentences with multiple possible interpretations, recognising

recurrent patterns in the type of content preceding and following IEF, and noticing when a input sentence is part of a multi-sentence quote. Text from further away in an article may provide helpful information about the topic, style and tone of the article, and may help systems recognise whether the input sentence is an outlier compared to the content and tone of other sentences from the same article.

2. **Event context**. News articles that cover a newsworthy event are rarely the only news article reporting that incident. Having access to a number of these other articles on the same event may help to notice when an article takes a unique angle to a topic or mentions information that is absent from other articles.

3. **Domain context**. The term domain commonly refers to genres of texts such as "news" or "science-fiction", but can also be used to denote the distribution over language of a specific corpus, or a subset of texts tied to a specific NLP task (Gururangan et al., 2020). There are many possible benefits to maximizing exposure of an IEF system to texts with the same distribution or a similar distribution to a given input sentence. Other articles from news outlets in general can provide information regarding typical journalistic strategies for framing entities without attracting accusations of bias (e.g. the use of third-party speakers to introduce opinions (Niculae et al., 2015)). Other articles from the same news outlets as the input data can help systems recognise idiosyncrasies of news outlets or differences between news outlets along partisan lines (e.g. the implications of the use of the terms "pro-life" versus "pro-choice"). Other news articles from political news can provide knowledge of domain-specific collocations and their connotations (e.g. "leading in the polls" or "declined to comment'). Finally, other news articles from the same time frame can provide additional information about the entities and topics that are most prominent in the input data.

Because the types of context described above vary considerably in terms of their scale, modeling each of them with the same technique is impractical. Instead, we use separate techniques for integrating each context type, each aimed to compress the additional information in a manner that is effective and efficient given the span of text involved. We do not consider the resulting approaches direct competitors of one another. Instead, we treat context types as additional sources of information to add to a representation of the target sentence and report which techniques provided significant benefits.

5.4.1 Context from the Same Article

Context from the article a sentence appears in may a) be helpful for interpreting the meaning of ambiguously worded sentences, and may b) relate input sentences to articles' general topic, style and tone. To address these distinct benefits of article context separately, we propose a technique for modeling direct textual context (neighbouring sentences) and a technique for modeling article context in the form of entire articles (whole article context).

5.4.1.1 Neighbouring sentences

Contextualising input sentences with their immediate surroundings requires the simultaneous modeling of sequences of sentences. Previous work has shown that this can be achieved with hierarchical sequence encoders (Dernoncourt and Lee, 2017; Jin and Szolovits, 2018; Papalampidi et al., 2019) and with a BERT adaptation that allows the joint labeling of multiple sentences with access to BERT representations of all the tokens in all sentences (Cohan et al., 2019). Of these, the latter compares most directly to our sentence-only baseline. This is desirable, because it lowers the chance that improvements in performance are due to differences in the approach other than the inclusion of neighbouring sentences.

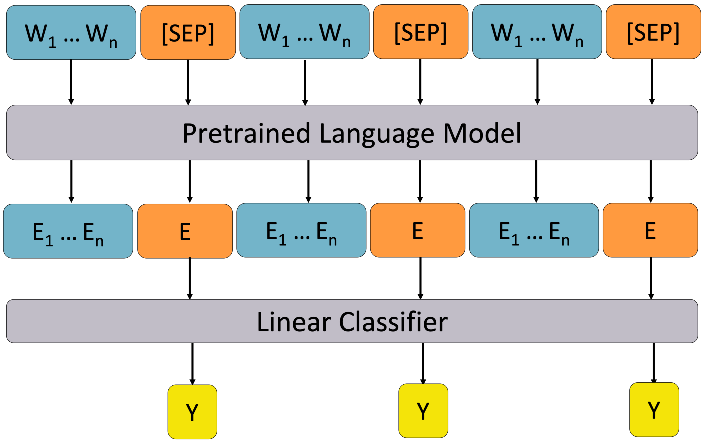

Fig. 5.3 Diagram illustrating Sequential Sentence Classification as proposed by Cohan et al. (2019): the separator tokens are used as contextualised sentence representations by the classifier layer to generate labels for each sentence in the sequence.

Cohan et al. (2019)'s BERT adaptation is called **Sequential Sentence Classification (SSC)**. As illustrated in Figure 5.3, it takes multiple sentences as its input sequence, generates embeddings for the separator tokens in the sequence, and classifies these embeddings with a linear layer that outputs as many labels as there are sentences in the input sequence.

SSC is designed for application to scientific abstracts, which are considerably shorter in length than news articles. To make the method applicable to larger documents, we introduce

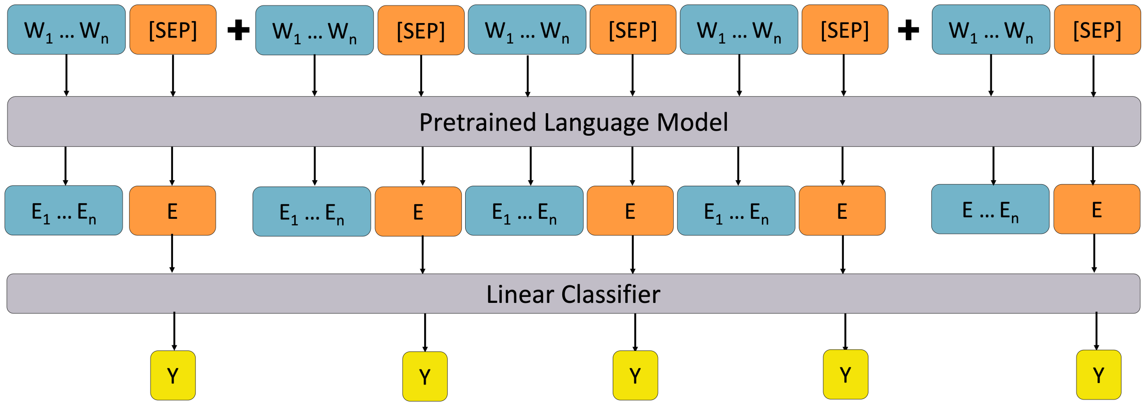

Fig. 5.4 Diagram illustrating Windowed Sequential Sentence Classification: sentence sequences are padded with the last sentence from the previous sequence and the first sentence of the subsequent sequence to prevent loss of context around sequence edges.

Windowed Sequential Sentence Classification (WinSSC). As illustrated in Figure 5.4, WinSSC also classifies sentences in sequence, but book-ends sequences with the last sentence from the previous sequence and the first sentence of the next sequence. The labels generated for the SEP-tokens of these book-ends are ignored during evaluation, but their inclusion in the training process ensures that each sentence in the sequence has context at both ends, thus mitigating loss of information along the edges of sequences. This is important when segmenting news articles as they tend to be long enough to require segmentation into several consecutive sections. The alternative (increasing sequence length to encompass entire articles) would generate inputs that are too large to be processed in memory by most training set-ups.

The BASIL data set does not maintain paragraph structures from the scraped news articles, so we set sequence lengths by experimenting with different candidate values. We experiment with sequence lengths of 5 and 10 sentences to assess the effect of changing section sizes, and compare our WinSSC method to the SSC method from Cohan et al. (2019) and to the best-performing baseline classifier, RoBERTa. We use code and hyperparameters from Cohan et al. (2019), but use BERT instead of SCIBERT weights. We report performance scores averaged across runs with 5 different seeds.

We find that performance not only does not increase, but even decreases when sentence representations are contextualised with either SSC or WinSSC (Table 5.6). Increasing the length of the sequence from 5 to 10 does not improve performance for either the SSC or the WinSSC model (F1 $= 38.19$ to F1 $= 38.22$ for SSC and F1 $= 38.67$ to F1 $= 37.44$ for WinSSC).

It is possible that data sparsity is at fault here. Our baseline results in Section 5.3.0.1 suggest that the BASIL corpus' size is not optimal for automatic classification. In the baseline

Model	Precision	Recall	F1-score
RoBERTa	43.12 ± 1.03	41.29 ± 1.37	42.16 ± 0.30
SSC-5	41.90 ± 1.00	36.16 ± 1.13	38.19 ± 0.98
SSC-10	43.84 ± 1.64	34.88 ± 0.71	38.22 ± 1.11
WinSSC-5	42.28 ± 0.99	36.94 ± 0.88	38.67 ± 0.82
WinSSC-10	43.20 ± 1.37	35.12 ± 2.41	37.44 ± 0.79

Table 5.6 Results of integrating direct textual context context with a Sequential Sentence Classifier without a window (SSC) or with a window (WinSSC) and a maximum sequence length of either 5 or 10.

scenario, the number of input sentences is around $6,400$ per fold. When performing 10-fold cross-validation for SSC with the maximum sequence length set to 5, the number of training sequences is instead around 1654 per iteration. With the maximum sequence length set to 10, this drops further to 856 sequences. This may be too small a number of sequences for the models to reliably learn interdependencies between sentences in a sequence from.

5.4.1.2 Whole Article Context

The key to whole article context integration is efficient modeling of long-distance relationships. Neural networks with recurrency have been shown to be a good technique for modeling relationships which exceed the immediate context (Mikolov et al., 2011; Hochreiter and Schmidhuber, 1997). Previous work has already used BiLSTMs (Hochreiter and Schmidhuber, 1997) to encode news texts in their entirety (Papalampidi et al., 2019). Inspired by the BiLSTM-based Context-Aware Model from Papalampidi et al. (2019), we propose a Context-Inclusive Model (CIM) which performs classification based on encodings of the target sentence and of sentences from some pre-determined context.

The CIM uses as input sentence representations from a fine-tuned pre-trained language model obtained by taking the average of the last four layers (Stage one of Figure 5.5). We also experimented with USE embeddings (Cer et al., 2018) and Sentence-Bert embeddings (Reimers and Gurevych, 2019) as sentence representations and found averaging the last four layers of a Transformer-based language model to be more effective. Sentence representations from a pre-defined context consisting of one or more context documents are processed by a BiLSTM to obtain one or more context representation. Context representations are concatenated with the representation of the target sentences, and the result is passed to a linear classifier to obtain a sentence label (Stage two of Figure 5.5).

The **Article Context-Inclusive Model (ArtCIM)**, the version of CIM illustrated in Figure 5.5, uses the average of the last four layers of a RoBERTa model that was fine-tuned

Model	Precision	Recall	F1-score
RoBERTa	43.12 ± 1.03	41.29 ± 1.37	42.16 ± 0.30
ArtCIM	38.81 ± 0.93	47.78 ± 1.82	42.80 ± 0.55

Table 5.7 Results of modeling article context with ArtCIM compared to a sentence-only baseline.

on the BASIL task to obtain sentence representations, and defines the context as all sentences from the same article that the target sentence appears in. ArtCIM's predictions are thus based on both the target sentence, which the baseline captures well, and on the news article the input sentence appears in, which the baseline has no access to. For details on the hyperparameters, which were chosen based on comparisons on held-out validation data, see Appendix C.

Table 5.7 shows that ArtCIM achieves much better recall than the baseline, but worse precision. Although its F1-score is slightly higher (42.80 compared to 42.16), this difference is not significant. It appears the additional context from the rest of the news article provides additional cues that help the model retrieve instances of framing. Those same additional cues, however, may also contain irrelevant information that, unfortunately, harms precision.

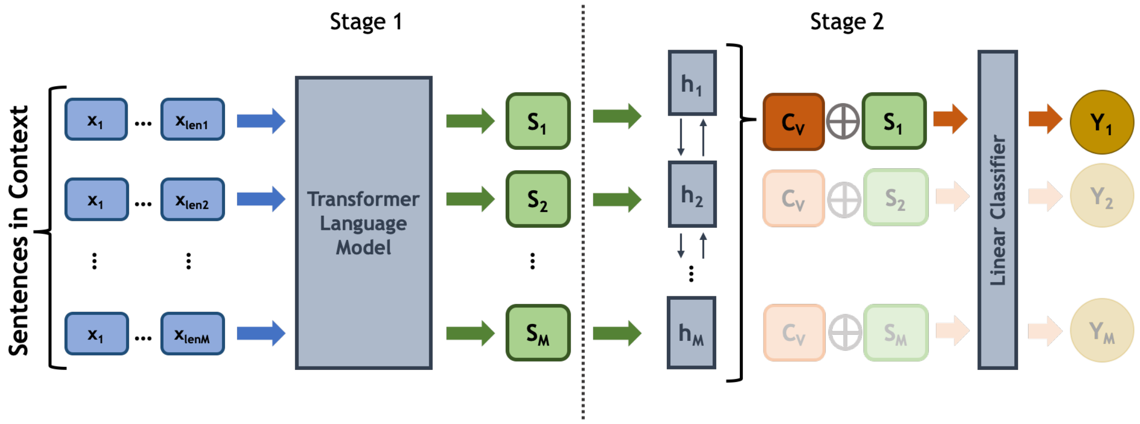

Fig. 5.5 Diagram of the Article Context-Inclusive Model. In Stage 1, sentence embeddings are obtained by encoding padded word sequences using a pre-trained language model. In Stage 2, the sequence of sentence representations from the same news article as the target sentence is encoded by a BiLSTM, and concatenated with the target sentence representation. The result of the concatenation is classified with a linear classifier to obtain a sentence-level prediction.

5.4.2 Context from Articles on the Same Event

The intended benefit of modeling event context is to model how a candidate sentence relates to coverage of topic, and to allow the IEF detection system to discover patterns that improve IEF predictions. For example, access to representations of related articles will reveal whether target sentences or article are highly similar to other sources' coverage of the same topic, or whether they are outliers. If being an outlier is associated with a higher likelihood of IEF, this information may boost performance. Another way in which event context may boost performance is by revealing event-specific patterns in reporting. For example, certain high-stakes events or high-profile events may be more controversial and attract more evaluative reporting, which may be associated with a higher chance of IEF in non-evaluative text.

To test whether representations of articles on the same event provide useful context for IEF detection, we train and test a version of CIM where the context consists of multiple context documents. Where ArtCIM contains a single BiLSTM, the **Event Context-Inclusive Model (EvCIM)** contains three BiLSTMs corresponding to the FOX, NYT and HPO items in BASIL event triplets (Figure 5.6). The linear classifier receives as input a concatenation of three context representations and the target sentence representation. We re-used the hyperparameters for ArtCIM, which were chosen based on comparisons on validation data (see C).

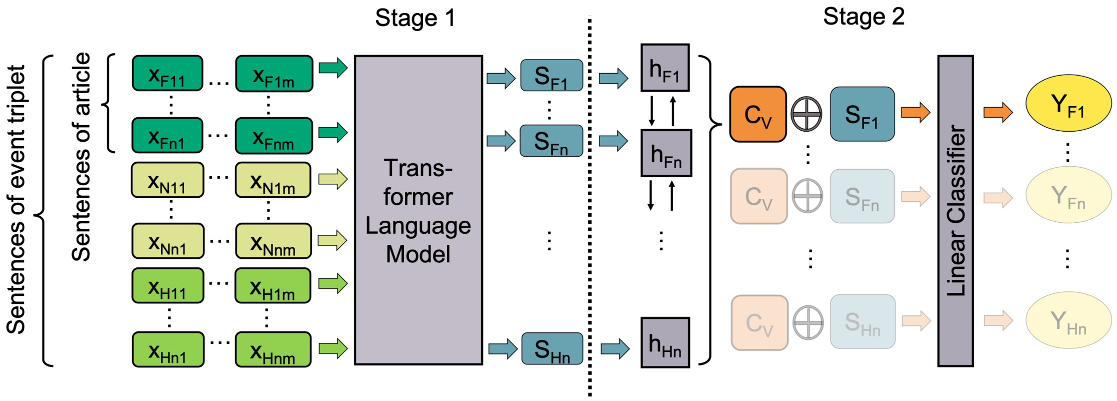

Fig. 5.6 Diagram of the Event-Context-Inclusive Model. In Stage 1, sentence embeddings are obtained by encoding padded word sequences using a pre-trained language model. In Stage 2, sequences of sentence representations from three articles on the same event are encoded by an equal number of BiLSTMs (only one shown in diagram).

Compared to the sentence-only baseline, EvCIM performs better in terms of recall. Unlike in ArtCIM's case, the decrease in precision compared to the baseline is smaller, and the overall F1-score is significantly better (44.10 vs 42.16, $p < .001$) (Table 5.8). To confirm that this boost in performance holds even when other base embeddings are used, we repeat

Model	**Precision**	**Recall**	**F1-score**
RoBERTa	43.12 ± 1.03	41.29 ± 1.37	42.16 ± 0.30
EvCIM	39.72 ± 0.59	49.60 ± 1.20	44.10 ± 0.15 †

Table 5.8 Results of incorporating event context with a Context-Inclusive Model compared to a RoBERTa sentence-only baseline. A dagger indicates that the improvement over the baseline is significant ($p < 0.05$).

Model	**Precision**	**Recall**	**F1-score**
BERT	40.54 ± 0.62	31.43 ± 1.56	35.39 ± 1.02
EvCIM	36.34 ± 1.18	40.48 ± 1.92	38.25 ± 0.52 †

Table 5.9 Results of incorporating event context with a Context-Inclusive Model compared to a BERT sentence-only baseline.

the experiment using BERT-embeddings rather than RoBERTa-embeddings as the basis of the model and as the basis of comparison. Table 5.9 shows that a BERT-based EvCIM model also significantly outperforms a BERT-based sentence-only classifier ($p < .001$) .

5.4.3 Context from the Same Domain

As discussed in Sections 2.1 and 2.1.2, media framing possesses a property of repetitiveness. Framing can be understood as a process of repeated combining of certain aspects of a topic instead of other equally applicable aspects, which establishes a common way of approaching that topic that influences public discourse. Sections 2.1 and 2.1.2 also state that while some frames may be universal, many of them are culturally specific and/or issue-specific. To illustrate this, a mention of professional experience in reality TV may be strongly indicative of negative framing if this mention appears near the name Trump in a left-leaning outlet around the time of the 2016 U.S. election, but not or less so when it co-occurs with other names, when it appears in a right-leaning outlet, or when it appears a decade earlier or later.

Because an instance of framing is likely to be part of a larger pattern that spans across a group of texts with shared properties, we suspect that IEF detection systems will benefit from incorporation of texts on other events but with those shared properties. We first pursue the computationally inexpensive and intuitive strategy of using source labels from the BASIL corpus as news outlet context (Section 5.4.3.1). We then explore the promising strategy of extended pre-training, where unlabeled text that resembles the training corpus is used to refine pre-trained embeddings (Section 5.4.3.2).

Model	**Precision**	**Recall**	**F1-score**
ArtCIM	38.81 ± 0.93	47.78 ± 1.82	42.80 ± 0.55
ArtCIM+Source	39.08 ± 0.53	45.18 ± 1.25	42.31 ± 0.63
EvCIM	39.72 ± 0.59	49.60 ± 1.20	44.10 ± 0.15
EvCIM+Source	39.76 ± 1.50	46.88 ± 2.42	42.96 ± 0.34

Table 5.10 Results of enhancing a Context-Inclusive Model with or without news source as an added feature.

5.4.3.1 Integration through source labels

A plausible factor in variation in entity framing is the source of news articles. News outlets may have idiosyncratic patterns of describing entities that reflect their values or their company culture. For example, a right-leaning news outlet may use words describing affiliation with left-leaning institutes to frame entities negatively, whereas those same words appear less frequently and/or are not associated with negative stance in other news outlets.

To investigate the impact of integrating news outlet context, we propose concatenating a representation of the source of an article as a feature in the CIM model at Stage 2 in Figures 5.5 and 5.6. We test this approach on the BASIL corpus by concatenating three-dimensional one-hot encodings representing either a FOX, NYT or HPO label with the input sentence embedding and the BiLSTM-based embedding of either the context article (**ArtCIM+Source**) or the event article triplet (**EvCIM+Source**).

Contrary to our expectations, while this approach improves precision, it reduces recall and lowers the F1-score compared to ArtCIM and EvCIM baseline (Table 5.10). As mentioned in Section 5.4.1.1, the BASIL corpus' size is not ideal for automatic classification, and it is possible that 100 instances of each source are too few to learn source-specific patterns from.

5.4.3.2 Integration through extended pre-training

A more involved but also more promising approach is to add domain-related information not as a feature, but through adaptation of word representations to the target domain. Previous work has shown that this can be accomplished with extended pre-training of existing generic language models. Extended pre-training in the form of both domain-adaptation (adaptation to a genre like "news" or "chemistry papers") and task-adaptation (adaptation to unlabeled text data sampled the same way as the training data) has been shown to improve downstream performance on classification tasks, without requiring computationally costly training of a domain-specific language model like BioBERT (Gururangan et al., 2020; Lee et al., 2020).

Extended pre-training is a promising approach for IEF detection, because it does not rely on labeled data. IEF detection is an extremely low-resource task with less than 8000 labeled sentences available to train on, which limits the potential of fine-tuning techniques like that proposed in Section 5.4.3.1. Creating additional IEF labels is very costly, because the annotation process requires attentive reading of news articles followed by discussion amongst annotators (Fan et al., 2019). Collecting unlabeled instances for pre-training, on the other hand, is not particularly challenging. Unlabeled news items may be gathered from existing corpora (Horne et al., 2018; Shu et al., 2018; Chen et al., 2018; Rose et al., 2002) or sampled from massive online web crawling resources.

As stated, positive results have been observed for extended pre-training in the form of both domain-adaptation and task-adaptation. Gururangan et al. (2020) demonstrated the effectiveness of domain-adaptation to the domain "news" by performing additional pre-training on news data from the RealNews dataset[2]. Although the improvement was less pronounced for the news task than other tasks, they observed improved performance for a downstream hyperpartisan news detection task. Performance improvements have also been shown for task-adaption, which is extended pre-training on data the task-specific corpus was also sampled from (Gururangan et al., 2020; Howard and Ruder, 2018). For optimal results, it is recommended to perform curated task-adaptation. In curated task-adaptation, collected additional data is filtered to ensure that it resembles the task corpus, typically by measuring similarity between candidate data points and the task corpus (Gururangan et al., 2020).

We test the hypothesis that extended pre-training encodes performance-enhancing domain-specific information for IEF detection by sampling and performing extended pre-training on unlabeled news data from the same source and time frame as the labeled BASIL training data. For our task-adaptation settings, we additionally pre-train RoBERTa on unlabeled BASIL sentences (**TAPT**) and we additionally pre-train (Gururangan et al., 2020)'s RealNews-adapted model on unlabeled BASIL sentences (**DAPT-TAPT**). For our curated task-adaptation setting, we perform additional pre-training on large corpora of curated BASIL-resembling data (**50NN-CurTAPT**, **150NN-CurTAPT** and **500NN-CurTAPT**). See Table 5.11 for the sizes of all pre-training datasets.

We create curated BASIL-resembling data as follows. First, we sample a large subset of Fox News, New York Times and Huffington Post articles from each year of the 2010's from the CommonCrawl News dataset[3]. The resulting sample consists of 1.35 million Fox News sentences, 2.33 million New York Times sentences and 675k Huffington Post sentences. Secondly, we curate our sample by mapping sentences to a feature space as Sentence-BERT

[2] https://paperswithcode.com/dataset/realnews

[3] https://commoncrawl.org/2016/10/news-dataset-available/

Dataset	**#Docs**
DAPT	25M
TAPT	250
DAPT-TAPT	25M
50NN-CurTAPT	226K
150NN-CurTAPT	479K
500NN-CurTAPT	924K

Table 5.11 Datasets used for extended pre-training: generic news domain data (DAPT), unlabeled BASIL data (TAPT), a combination of the two (DAPT-TAPT), and a selection of documents (where each document is a sentence) from the same source and time frame and with high similarity to the BASIL corpus (CurTAPT).

embeddings (Reimers and Gurevych, 2019), computing their cosine similarity to similarly encoded BASIL corpus sentences, and selecting for each embedded BASIL sentence a k-sized set of nearest neigbours from the embedded sample sentence. Following Gururangan et al. (2020) we construct subsets of the 50, 150 or 500 nearest neighbours (50NN-CurTAPT, 150NN-CurTAPT and 500NN-CurTAPT). These settings correspond to sets of 226k, 479k and 924k sentences respectively.

To compare performance, we perform additional pre-training of RoBERTa on the datasets described above and fine-tune the resulting adapted models on labeled BASIL data to create IEF classifiers. For the pre-training phase, we use code made available by Gururangan et al. (2020) with its default hyperparameters.

Table 5.12 shows that the DAPT and combined DAPT-TAPT settings fail to improve over the baseline, as do the 50NN and 150NN CurTAPT settings. The 500NN CurTAPT setting performs only as well as the baseline. The TAPT setting which has seen only unlabeled BASIL sentences in addition to the initial RoBERTa pre-training data performs best and slightly outperforms the baseline, but not significantly.

These results are quite surprising given the effectiveness of similar approaches in other work. Our suspicion is that, first of all, much of the relevant trends in news tasks are already captured in RoBERTa's initial pre-training input, which contains a substantial amount of news texts. This echoes the finding in Gururangan et al. (2020) that domain-adaptation and domain-and-task-adaptation are not as helpful in the news domain as in other domains. Secondly, we suspect that the task of predicting IEF labels on held-out BASIL data is exceedingly difficult for reasons unrelated to the quality of the word embeddings that are used, but that instead pertain to perhaps not only simply the limited size of the labeled data, but also the labeling strategy used to create gold labels.

Model	**Precision**	**Recall**	**F1-score**
RoBERTa	43.12 ± 1.03	41.29 ± 1.37	42.16 ± 0.30
DAPT	46.87 ± 1.32	36.45 ± 1.27	40.97 ± 0.32
TAPT	46.49 ± 1.74	40.28 ± 1.91	43.12 ± 1.07
DAPT-TAPT	46.69 ± 1.50	37.41 ± 2.37	41.47 ± 1.00
50NN-CurTAPT	43.80 ± 1.21	40.11 ± 1.30	41.87 ± 1.21
150NN-CurTAPT	43.95 ± 1.36	37.41 ± 2.13	42.01 ± 1.18
500NN-CurTAPT	44.08 ± 1.09	40.59 ± 0.93	42.24 ± 0.25

Table 5.12 Results on IEF classification of BASIL data after adapting RoBERTa embeddings to generic news domain data (DAPT), unlabeled BASIL data (TAPT), a combination of the two (DAPT-TAPT), and to a selection of unlabeled sentences from the same source and time frame and with high similarity to the BASIL corpus (CurTAPT).

5.4.4 Conclusion

We explore the impact of incorporating three kinds of context in IEF detection: article context, event context and domain context. Integrating article context from the direct vicinity of the target sentence by training a neural network on sequences of consecutive sentences proved unhelpul for IEF detection, as did integrating the entire target article through recurrent encoding of articles' sentence embeddings. What did yield significant improvements over the strong sentence-only baseline was to incorporate context from other articles on the same event. This model, the EvCIM model, classifies concatenations of same-event articles' encodings and input sentence embeddings. We also experimented with incorporating domain context by either treating news outlet labels as a classification feature, or by adapting RoBERTa embeddings to various kinds of domain-specific data. Neither approach proved successful, meaning that EvCIM outperformed sentence-only classification with either vanilla, domain-adapted and task-adapted RoBERTa embeddings. These findings suggest that context from other articles on the same event is particularly valuable for IEF detection.

Chapter 6

Performance Analysis of Context-Inclusive Entity Framing Detection

6.1 Introduction

In the previous chapter, we explored various strategies for context-inclusion in informational entity framing detection in news articles. The most successful strategy was to provide a sentence classifier with representations of articles covering the same event as the target article. This Event Context-Inclusive Model (EvCIM) significantly outperformed both a BERT and a RoBERTa-based sentence-only baseline. In this chapter, we provide evidence that EvCIM's success is attributable to access to context, instead of to, for example, differences in the training procedure or model architecture.

To establish a relationship between EvCIM's success and access to context, we perform comparisons between EvCIM and the best-performing sentence-only model (RoBERTa) on BASIL instances that may be difficult to classify without access to context. No gold labels exist for the degree of context-dependence of IEF-containing sentences, so we instead analyse the relative improvement in performance of EvCIM over a sentence-only baseline using proxies for context-dependence. Our proxies for context-dependence are properties of instances of which one can reasonably expect that they increase instances' difficulty. For example, one can reasonably expect that it is easier to determine whether a sentence contains IEF if it contains subjective language that suggests the writer intended the text to have an evaluative purpose. We perform a suite of performance tests where we compare groups of instances which vary in expected difficulty and report the relative increase of EvCIM over

RobERTa, and find that EvCIM performs as one would expect if it leveraged context on all but one of the tests.

We examine seven dimensions of difficulty. On the sentence-level, we measure the impact of sentence length as well as of the presence of subjective language in the sentence. We also assess the impact of the commonness of the target of IEF spans, of their polarity and directness, and of whether they are part of a quotation. On the article-level, we compare performance on articles with different political leanings (left, center and right).

For each of the seven properties, we take a stance on how we believe they impact classification difficulty, and then test the hypothesis that EvCIM outperforms a sentence-only baseline by a higher margin on the class with the higher difficulty. The next paragraph lists the properties and their predicted impact on difficulty along with our reasoning for that prediction illustrated by an example from Table 6.1.

1. **Sentence length**: we posit that longer instances are harder to classify out of context than shorter instances because they contain more complex messages (1.2 vs 1.3 in Table 6.1).

2. **Subjective language**: we posit that instances without subjective language are harder to classify out of context than instances which do contain subjective language, because their evaluative purpose must be inferred and is not hinted at by the language that is used (1.3 vs 3.3, which contains the word “complained”, in Table 6.1).

3. **Target**: we posit that instances of IEF whose targets are not frequently discussed are harder to classify than instances with IEF targets that are very frequently discussed, because models cannot rely on recurring patterns of speech in the corpus when labeling them (1.2 vs 2.2 in Table 6.1).

4. **Polarity**: we posit that positive IEF is harder to recognize than negative IEF because negative news frames are more common than positive ones and therefore more likely to be encoded in word embeddings from pre-trained language models (Soroka and McAdams, 2015) (1.3 vs 2.3 in Table 6.1).

5. **Direction**: we posit that indirect IEF is harder to classify out of context than direct IEF because more steps of inference are required to recognise it (3.3 vs 3.2 in Table 6.1).

6. **Part of quotation**: we posit that IEF is harder to recognise outside of quotes than inside quotes, because of the proven relationship between quotations and framing in news (Niculae et al., 2015; Fan et al., 2019) (3.3 vs 3.2 in Table 6.1).

Idx	**Sentence**	**Source**	**IEF**	**Tar**	**Pol**	**Dir**
1.1	Former Arkansas Gov. Mike Huckabee announced Tuesday he is running for president [. . .].	FOX	0	N/A	N/A	N/A
1.2	Mr. Huckabee opposes same-sex marriage, suggesting as recently as February that homosexuality is a lifestyle choice akin to drinking or swearing.	NYT	1	0	0	1
1.3	He was the longest-serving Arkansas governor, from 1996 to 2007.	HPO	1	0	1	1
2.1	“Trump says he wants to run the nation like he’s run his business,” Mr. Bloomberg said on Wednesday.	NYT	0	N/A	N/A	N/A
2.2	Bloomberg contrasted his business history with Trump’s, saying "I’ve built a business, and I didn’t start it with a million-dollar check from my father."	FOX	1	1	0	1
2.3	“But Trump’s business plan is a disaster in the making."	HPO	1	1	0	1
3.1	The states of Nebraska and Oklahoma filed a federal lawsuit in the U.S. Supreme Court Thursday [. . .].	HPO	0	N/A	N/A	N/A
3.2	"While Colorado reaps millions from the sale of pot, Nebraska taxpayers have to bear the cost."	FOX	1	0	0	1
3.3	[Sheriff Adam Hayward of Deuel County, Neb.] has complained that marijuana arrests have strained his jail budget.	NYT	1	0	0	0

Table 6.1 Example instances from three news stories in the BASIL corpus, along with their news source and labels for IEF (contains IEF: 1, does not contain IEF: 0), target (common: 1, uncommon: 0), polarity (positive: 1, negative: 0) and direction (direct: 1, indirect: 0). Note that Examples 2.3 and 3.2 contain sentiment-bearing language but are nevertheless considered instances of IEF because the opinion is contained within a direct quote.

7. **Article leaning**: we posit that instances from an article with a centrist political leaning are harder to classify than instances from a recognizably left-leaning or right-leaning source, because the language in centrist articles is more neutral (1.2 vs 3.3 in Table 6.1).

6.2 Sentence Length

By virtue of containing more clauses and more meaning-bearing tokens, longer sentences convey more information on average than shorter ones. We predict that this additional information increases the complexity of a sentence, and therefore increases the difficulty of determining whether all or some of it constitutes a case of IEF. Subtle positive or negative connotations of a subset of tokens, may be weakened or negated in other parts of the sentence, resulting in an overall sentence-level embedding from which the relevant information pertaining to framing can nolonger be recovered. For example, the phrase "Trump's business plan is a disaster in the making" in Table 6.1 clearly reflects negatively on the entity Donald Trump. This is less obviously the case if we lengthen the sentence by adding a phrase which introduces a possible motive for Bloomberg's criticism, as in: "Bloomberg, Trump's competitor and long-time rival, was quick to call Trump's business plan a disaster in the making".

We test the impact of sentence length on performance difference between a context-inclusive versus a sentence-only model by comparing EvCIM and RoBERTa's performance on BASIL sentences of varying lengths. We measure sentence length as the number of tokens as determined by an off-the-shelf tokenizer[1]. We partition the data into bins corresponding to quartiles of the length metric, and report the relative increase in F1-score of EvCIM compared to RoBERTa.

As predicted, EvCIM outperforms the baseline by a higher margin on longer sentences (Table 6.2). The difference in performance between EvCIM and RoBERTa is not significant on the shorter sentences, i.e. on the first and second quartile, but reaches significance for the third and fourth quartile (4.89% and 7.91% relative performance increase).

6.3 Subjective Language

We suspect that the presence of subjective language facilitates IEF classification, because subjective language can be used as an indicator of IEF, in the absence of which IEF must be inferred from either the semantics of the sentence, or from its context. A common source

[1] SpaCy's en-core-web-md, version 3.1.0

Q	N	%Bias	RoBERTa	EvCIM	%Increase
38-110	1893	18.96	44.48	48.00†	7.91
28-37	1961	15.81	40.02	41.98†	4.89
20-27	1911	13.40	41.01	41.10	0.22
0-19	2212	13.38	43.82	44.02	0.46
All	7977	15.31	42.48	44.10†	3.81

Table 6.2 RoBERTa and EvCIM performance (F1) and EvCIM's relative improvement over the baseline by sentence length (quartiles). The first column shows the number of tokens per sentence in each quartile (Q4 to Q1), the second the number of sentences in each quartile and the third the percentage of biased sentences per quartile. The final column shows the relative improvement of EvCIM over the baseline. A dagger indicates that the improvement is significant ($p < 0.05$).

of subjective language in IEF spans is the presence of third-party opinions, which can be presented through a direct quote, but they can also be expressed indirectly such as in the case of the verb "complain" in Instance 3.3 of Table 6.1. Subjective language may also appear in phrases describing the positive or negative consequences of events, such as "strained his budget" in 3.3 in Table 6.1. We predict that in all these cases, the subjective language can be used as a heuristic for predicting IEF that decreases the need for context and shrinks the performance gap between RoBERTa and EvCIM.

We choose not to use the BASIL corpus annotations of lexical bias (LB) to assess the impact of the presence of subjective language on EvCIM gains. This is because the distribution of BASIL's LB labels has some surprising characteristics. LB labels are quite rare, with only 448 out of 7977 BASIL sentences containing LB. They also do not tend to co-occur with IEF as one might expect: IEF is less likely to co-occur with LB in the same sentence (9.82%) than it is to appear alone (15.63%). We suspect that this is due to the method of annotation. BASIL annotators labeled a sentence as containing LB by judging whether their opinion of an entity was being swayed by the choice of words, which means they would not have labeled word choices that are subjective, but had no perceivable effect on their opinion.

Instead of using LB labels to assess the impact of the presence of subjective language on EvCIM gains, we use a quantitative indicator for the presence of subjective language based on the MPQA Subjectivity Lexicon (Wilson et al., 2005). We label each instance for whether it contains at least one strongly subjective item from this lexicon. According to this definition, 2363 instances out of 7977 contain subjective language, and IEF is more likely to occur in sentences with subjectivity (18.92%) than sentences without subjectivity (13.74%).

Subjectivity	N	%Bias	RoBERTa	EvCIM	%Increase
No	5614	13.68	40.19	41.97†	4.44
Yes	2363	19.17	46.04	47.41†	2.96
All	7977	15.31	42.48	44.10†	3.81

Table 6.3 Performance (F1) and EvCIM's relative improvement over the baseline on items that do or do not contain at least one strongly subjective token from the MPQA Subjectivity Lexicon.

The performance test for subjective language reveals a significant increase in performance of EvCIM over RoBERTa for both groups (Table 6.3, $p < 0.05$), but as predicted, the increase in performance is higher for sentences that contain no subjective language (4.44%) than for sentences that do (2.96%).

6.4 Target

Some entities are discussed more frequently in the news than others. For frequently discussed entities, recurring patterns and turns of phrases may emerge in discourse that statistical models can commit to memory. When classifying text about the actions of lesser-known entities, statistical models are then at a comparative disadvantage. Any benefit from having had access to context during training may be more apparent in the latter case.

We test this hypothesis by separating model performance by the commonness of IEF-containing sentences' targets. We determine the commonness of a target using BASIL's main entities annotations, which are lists of the entities that are most prominently featured in a given article. Main entities can be a named entity like "Barack Obama" or a group of people like "Republican Lawmakers". For each target of an IEF-containing sentence, we count how frequently it appears as an article's main entity. We rank entities by these frequencies and label the top ten as common entities and the rest as uncommon entities (see Appendix D for the full list). The top ten consists of: Donald Trump (85 articles), Barack Obama (54 articles), Hillary Clinton (24 articles), Republican Lawmakers (20 articles), Democratic Lawmakers (18 articles), House Democrats (7 articles), Paul Ryan (7 articles), Nancy Pelosi (6 articles), Michael Flynn (6 articles) and Robert Mueller (6 articles).

We compare how RoBERTa and EvCIM perform in terms of recall for instances with common or uncommon targets[2]. We also predict that articles that feature common entities are as a whole easier to classify than articles that do not, and therefore also analyse performance

[2] F1-scores cannot be compared here: IEF target commonness is an attribute of IEF instances, not of sentences that do not contain IEF.

Common Target	N	RoBERTa	EvCIM	%Increase
No	719	38.75	49.01†	26.49
Yes	502	41.75	50.44†	20.80
All Biased	1221	39.98	49.60†	24.05

Table 6.4 Performance (recall) by commonness of the target entity.

Common Main	N	%Bias	RoBERTa	EvCIM	%Increase
No	3204	13.76	39.36	42.39†	7.71
Yes	4773	16.34	44.19	45.12†	2.10
All	7977	15.31	42.48	44.10†	3.81

Table 6.5 Performance (F1) by whether instances' article features at least one common entity.

in terms of F1-score by whether at least one of the main entities of the article is a common entity as defined above.

The commonness of an instances' target has the predicted effect on sentence-level recall (Table 6.4). If the target entity is not common, the relative increase in recall of EvCIM over the baseline is higher (26.49%) than if it is common (20.80%). We also find as predicted that EvCIM's improvement in F1-score is higher for instances from articles with at least one common main entities (7.71%) than for articles with only uncommon main entities (2.10%) (Table 6.5).

6.5 Polarity

Negative framing is more common than positive framing in the news (Soroka and McAdams, 2015). In the BASIL corpus, too, negative IEF is more common: 79% of all IEF spans reflect negatively on the target. Because the data contains fewer examples of positive IEF, we predict that the sentence-only baseline will perform worse on these items, whereas EvCIM will be able to resolve some of these cases using patterns it learned while combining sentence-level with article-level and event-level context.

This prediction is confirmed by the performance analysis. When measuring performance by polarity, we observe a large drop in performance for positive instances for the baseline in particular (from 43.18 to 28.11). We also observe that EvCIM's recall exceeds the baseline by a much higher margin for positive instances (45.88%) than for negative instances (20.22%) (Table 6.6).

Polarity	**N**	**RoBERTa**	**EvCIM**	**%Increase**
Positive	259	28.11	41.00†	45.88
Negative	962	43.18	51.91†	20.22
All Biased	1221	39.98	49.60†	24.05

Table 6.6 Performance (recall) by polarity.

Indirect IEF	**N**	**RoBERTa**	**EvCIM**	**%Increase**
Yes	152	42.24	49.61†	17.45
No	1069	39.66	49.60†	25.05
All Biased	1221	39.98	49.60†	24.05

Table 6.7 Performance (recall) by direction.

6.6 Direction

The direction or aim of IEF refers to whether the framing applies to the main entity directly, or indirectly through an intermediary. An example of indirect entity framing is the use of a phrase like "Obama Administration" that indirectly target the entity Barack Obama. Because indirect framing requires an additional step of inference, it is most likely harder to classify correctly, and it is possible that access to context helps resolve some cases that can not be recognised by a sentence-only baseline.

The performance analysis of F1-scores by IEF aim does not confirm this. EvCIM's recall does not exceed the baseline by a higher margin for instances of indirect IEF than for direct IEF (Table 6.7).

6.7 Part of Quotation

Because they allow journalists to present an opinion without taking a stance themselves, quoting patterns are a common means of introducing bias in news (Niculae et al., 2015), and are associated with IEF in the BASIL corpus, too (Fan et al., 2019). We predict that neural approaches detect this relationship and learn to associate quotation marks or other indicators of quoting to IEF. Instances that are not in quotes may thus be harder to recognise, making context-inclusion potentially more beneficial when classifying such instances. For this performance analysis we use existing BASIL quote annotations that specify for each instance of IEF whether it is part of a third-party quote or not.

Table 6.8 shows that both models have considerably better recall for IEF in quotes than IEF outside of quotes. Additionally, as predicted, the gains of EvCIM with respect to the baseline are decidedly higher on non-quotes (40.70%) than on quotes (15.07%).

In Quote	N	RoBERTa	EvCIM	%Increase
No	624	27.40	38.56†	40.70
Yes	597	53.13	61.14†	15.07
All Biased	1221	39.98	49.60†	24.05

Table 6.8 Performance (recall) by whether the IEF span exists inside or outside of a quote.

Outlet	Right	Center	Left	All
FOX	50	38	12	100
NYT	15	54	31	100
HPO	10	52	38	100
All	75	144	81	300

Table 6.9 Number of right-leaning, centrist and left-leaning articles from each news publisher.

6.8 Political Leaning

U.S. news analysts distinguish between centrist news outlets and outlets that are known to report on events from a noticeably left-leaning or right-leaning point-of-view[3]. We expect that on average, newspapers and articles with a less pronounced political leaning contain framing that is less overt in nature. We predict that, consequently, instances from centrist articles require more context to recognise, and that EvCIM will outperform the baseline by a higher margin for these instances.

We compare performance differences between centrist and non-centrist articles in two ways; separated by their outlet and separated by their political leaning annotation. The BASIL corpus contains articles from three outlets: Fox News, the New York Times, and the Huffington Post. According to Budak et al. (2016), these are a right-leaning, slightly left-leaning, and strongly left-leaning outlet respectively. The BASIL corpus also contains article-level labels for political leaning, with the options being "Right", "Center" and "Left". Table 6.9 shows the distribution of labels across these two dimensions for all 300 BASIL articles. Note that each outlet publishes a substantial amount of articles that were considered centrist by the human annotators.

If our prediction for the impact of leaning is correct, EvCIM should outperform RoBERTa most on the subset of 100 NYT articles, and on the 144 articles with label "Center". The performance tests indeed show a higher relative improvement for these categories (Table 6.10). There is a 6.03% increase in performance for instances from the NYT compared to 3.17% and 2.24% for the other outlets. The improvement for instances from articles with a centrist stance is 6.69%, compared to 1.07% and 2.34% for the other labels. Additionally, the difference for

[3] https://www.allsides.com/media-bias/media-bias-ratings

Outlet	N	%Bias	RoBERTa	EvCIM	%Increase
FOX	2633	15.65	45.85	47.31†	3.17
NYT	3048	14.93	40.57	43.01†	6.03
HPO	2296	15.42	40.90	41.82	2.24
All	7977	15.31	42.48	44.10†	3.81

Table 6.10 Performance (F1) by the news outlet instances are from.

Stance	N	%Bias	RoBERTa	EvCIM	%Increase
Right	2010	15.82	43.46	43.92	1.07
Center	3660	14.07	42.55	45.40†	6.69
Left	2307	16.82	41.61	42.58	2.34
All	7977	15.31	42.48	44.10†	3.81

Table 6.11 RoBERTa and EvCIM performance (F1) and EvCIM's relative improvement over the baseline by the stance of the instances' articles.

articles whose stance is centrist is significant ($p < 0.05$), whereas the difference for articles with a right-leaning or left-leaning stance label is not.

An interesting additional observation is that both models perform better on FOX articles than on NYT and HPO articles, with the difference ranging from 4.30 to 5.28. Fan et al. (2019) have stated that the three news sources included in the BASIL corpus differ in both the polarity and the targets of their instances of IEF. It is possible that RoBERTa and EvCIM capitalize on these and possibly other idiosyncrasies of FOX articles to make more accurate predictions for them.

6.9 Conclusion

Our error analysis of seven dimensions of classification difficulty shows that for six out of seven dimensions EvCIM performs better than a sentence-only model baseline on instances that can reasonably be expected to be harder to classify out of context. Specifically, we observed higher gains for longer sentences, sentences that do not contain subjective language, IEF spans with uncommon target entities and from articles with uncommon main entities, IEF spans with positive polarity, IEF spans that are not part of a quotation and finally, for sentences from politically centrist outlets and articles. The consistency with which EvCIM outperforms the baseline on these difficult cases supports the notion that EvCIM's success is not due to better memorization of the training data, but due to the learning of context-dependent patterns that boost performance in situations where the sentence on its own is not sufficiently informative.

Chapter 7

Conclusions and Future Work

Work on the framing of entities is very rare, but as our studies on naming and titling have shown, small changes in presentation are likely to impact attitudes towards entities, just like previous research have shown is the case for issues and events. We contribute evidence that in order to perform better automatic detection of framing towards entities in news, detection systems need access to representations of context in the form of related articles covering the same event.

Our evidence for a likely framing effect of small changes in presentation of entities is provided by our studies on naming form variation and title usage, which constitute the first quantitative work on this topic. We find that there exists a relation between mentioning or omitting part of an entities' name when referring to them, and of including or omitting their title, on the perceived attitude (stance) of a text towards that entity. Using a naming form that is more formal, which includes using a title, is positively correlated to the perceived stance of social media messages mentioning politicians. We show this to be the case for two languages and two titles on the first two existing corpora labeled for naming forms. Following observations from sociolinguistics about the status-signaling function of naming, this finding indicates that the choice of naming form signals, intentionally or unintentionally, the perceived status of the politician. This supports the view that names and titles, which are at face value objective properties of a person, carry a connotation that can influence the opinion-forming process in political discourse. We also show that this status-signaling function is weaker in left-leaning discourse, and that female German politicians are referred to by their academic title equally often as male politicians.

Our evidence that news articles on the same event are useful context information for Entity Framing detection comes from our experiments with Informational Entity Framing detection systems. We show that despite the difficulty of this detection task, including articles from the same event when training a model can improve detection of entity framing through tangential, speculative or background information. Our proposed technique for including same-event context, the Event Context-Inclusive Model, outperforms a strong baseline on an improved training/test split, establishing a new state-of-the-art for the task. We provide evidence that it is the access to context rather than other properties of the neural architecture that provide this boost in performance by demonstrating a higher gain in performance on difficult instances on six dimensions of difficulty.

Our study on the impact of German political entities' gender on the frequency with which Twitter users apply titles to them found no gender bias in title usage. This is surprising, and an unexplored question of relevance for future work is whether the titles used for male and female politicians are perhaps equally frequent but different in terms of the strength of their positive correlation to stance. A related open question is whether title usage and highly formal naming in general is ever sarcastic in intention, and whether sarcastic use plays a role in weakening the positive association between naming formality and stance in left-leaning discourse. Of less interest to the study of framing of political entities, but of broader sociolinguistic interest is whether the formal naming correlates positively with stance more generally, across domains and entity types, or whether it is typical of either political discourse or mentions of politicians.

In order to give more precise and reliable answers to questions regarding possible framing effects of naming, we encourage extensions to ETTC and GTTC annotations. An added layer of annotations that states whether naming forms are used in reference or terms of address would make it possible to refine our claims about the function of naming in political discourse. If enough forms exist for both categories, such annotations would make it possible to determine whether there is a difference in the strength or direction of the positive correlation between naming/titling and stance depending on whether the naming form is a reference form or a term of address. If one category is far more frequent than the other, the infrequent one can be eliminated, allowing a more precise and reliable claim regarding the stance relation of the more frequent one.

Another suggestion for additional labeling concerns the problem of whether the naming formality can be shown to not only coincide with possible stance, but to contribute to annotators' perception of the stance of a text. To know whether formal naming or title usage can enhance or even cause the perception of positive attitude towards entities, it is necessary to obtain stance labels for permutations of texts which vary only in the formality of their naming form. While exploring designs for controlled experiments during the early stages of planning our experiments, our impression was that is it very difficult to design a large enough number of prompts that allow variation of the political entity and the naming form without losing coherence. However, now that the ETTC and GTTC exist, future work may be able to use the provided instances and the proposed ranking of formality to gather additional stance labels for copies of ETTC or GTTC instances that have been altered only in terms of their naming form. This would make it possible to assess whether increasing the naming formality of instances also increases perceived stance.

When it comes to our study of the impact of context in IEF, although we discovered an effect of including event context that held up across performance tests, some of the negative outcomes of context-integration experiments were surprising. Particularly surprising was the finding that additional pre-training to tune word embedding to the exact domain of 2010-2020 U.S. political news did not improve a classifier using those word embeddings to produce sentence-level predictions for texts from that domain. Additional work testing the robustness of the baseline models may be needed in order to determine whether the baseline task is solid and valid. This can include comparing pre-trained language model (PLM) performance to SVM and NB classifiers and/or using data augmentation techniques like those presented in Dhole et al. (2021), to analyse PLM's robustness to permutations in the data.

In general, but especially if PLM baselines lack robustness, it may be helpful to reconceptualise the IEF detection task. Annotating for IEF is a highly subjective task and achieving agreement between annotators can require extensive discussion (Fan et al., 2019). If unify differing interpretations of texts into a single gold label is so effortful for humans, perhaps binary classification is an inappropriate approach to IEF detection. Future models should perhaps be designed to take as training input fixed-size arrays of annotations and to output distributions of votes over annotators (compare to Nie et al. (2020)). This would of course require that data set creators provide the raw annotations underlying their gold labels.

In addition to providing arrays of labels instead of a single gold label, future data sets may also want to use reader-anchored prompts like those described in our naming and titling experiments in their annotation procedure. Encouraging annotators to take the point-of-view of a supporter of an entity when determining whether text reflects positively or negatively on them may help reduce variation in annotators' response. Annotators' can also be asked to provide details regarding their background as a tool to select a variety of annotators with differing perspectives (Spinde et al., 2021). They can also be asked to provide explanations for their choice of label which can be used as meta-data to supervise training (Wiegreffe and Marasović, 2021).

Assuming a robust sentence-only baseline, there are a number of interesting avenues to take with regards to further experiments on context integration. Given the evidence that event-level context is valuable for IEF detection, future work may want to perform more advanced contrasting between same-event articles by enhancing models with modern attention mechanisms (Galassi et al., 2020), or by using similarity measures between sentences as features. Because news articles tend to feature high-level information at the beginning and tangential or background information towards the bottom, another avenue is to experiment with representations of the position of a target sentence, or representations of paragraphs of context articles rather than encodings of entire articles.

It is also possible that IEF should be viewed not in terms of the mentioning of aspects that others omit, but in term of repeatedly describing an entity in terms sources with a different agenda do not. We found our sentence-only baseline and our best-performing context-inclusive model both perform better on BASIL's Fox News instances than its New York Times and Huffington Post instances. Firstly, this suggests that generic IEF detection system need to be trained on a variety of sources that represent different groups in society in order to generalise to unseen sources. Secondly, it suggests high-quality IEF detection may require outlet-specific modeling. Given a large enough data set, it may be possible to train systems, for example through outlet-specific extended pre-training, that excel at detecting outlet-specific ways of framing entities.

We hope our contributions and the suggestions in this final section will inspire an increase in work on Entity Framing of various kinds.

Appendix A

List of News Accounts

List of news accounts filtered out from the GTTC (Section 4.2.1.2)

- ARD_BaB
- ARD_Presse
- ARDde
- AZ_Augsburg
- BERLINER_KURIER
- BILD
- BILD_Koeln
- BILD_LA
- BILD_Muenchen
- BILD_National11
- BILD_Nuernberg
- BILD_Politik
- BILD_Promis
- BILD_Ruhrgebiet
- BILD_S_Anhalt
- BILD_Stuttgart
- BILD_aktuell
- BILDamSONNTAG
- BMOnline_pol
- BR24
- BR_Niederbayern
- BR_Schwaben
- BR_kontrovers
- BR_quer
- Berlin_Ticker
- BlnTageszeitung
- BreakingIEN
- BremerhavenNew1
- China_Welt_News
- DACH_Politik
- DAZheute
- DLF
- DLFNachrichten
- DLF_Berlin
- DLF_Sport
- DLF_Umwelt
- DailySabahDE
- DasErste
- EpochTimesDE
- FAZ_Feuilleton
- FAZ_Finance
- FAZ_NET
- FAZ_Politik
- FAZ_RheinMain
- FAZ_Sport

- FAZ_Top
- FAZ_Vermischtes
- FAZ_Wirtschaft
- FOCUS_Eil
- FOCUS_Magazin
- FOCUS_TopNews
- FreiePresseNet
- GN_Nordhorn
- HAZ
- HNA_online
- HandelsblattGE
- HuffPostDE
- JuedischeOnline
- KNA_Redaktion
- KURIERat
- Kreiszeitung
- LVZ
- MDRAktuell
- MDR_SAN
- MDR_SN
- MDRpresse
- NDRinfo
- NDRpresse
- NDRreporter
- NDRsh
- NN_Online
- NZ_Online
- NewsAustria
- NewsDeutsch
- NewsFinder24
- NewsFrontDE
- Nordkurier
- PNN_de
- PolitikStandard
- RT_Deutsch
- SPIEGELNOLINE
- SPIEGELONLINE
- SPIEGELTV
- SPIEGEL_24
- SPIEGEL_EIL
- SPIEGEL_Kultur
- SPIEGEL_Pano
- SPIEGEL_Politik
- SPIEGEL_Sport
- SPIEGEL_Top
- SPIEGEL_Wirtsch
- SPIEGEL_Wissen
- SPIEGEL_alles
- SPIEGEL_live
- SPORT1
- SWPde
- SWR2
- SWRAktuell
- SWRAktuellBW
- SWRAktuellEil
- SWRAktuellRP
- SZ
- SZ_Bayern
- SZ_Digital
- SZ_Eilmeldungen
- SZ_Medien
- SZ_Panorama
- SZ_Politik
- SZ_Sport
- SZ_TopNews
- SZ_Wirtschaft
- SimplyNewsDe
- TAG24
- TAG24BI
- TAG24CH
- TAG24DD
- TAG24FFM
- TAG24HH

- TAG24Koeln
- TH24DeineNews
- Tag24B
- Tag24S
- Tagesspiegel
- Tageszeitung1
- WAZ_Redaktion
- WDR
- WDR2
- WDR_live
- WDRinvestigativ
- WELT_Kultur
- WELT_Politik
- WELT_Sport
- WELT_Wissen
- WELTnews
- Weltspiegel_ARD
- Weltwoche
- WienerZeitung
- WirtschaftCom
- ZDF
- ZDFhessen
- ZDFheute
- ZDFnrw
- ZDFsachsen
- aNewsDeutsch
- aktuelle_stunde
- ard_Warschau
- ardmoma
- badischezeitung
- berlinerzeitung
- bzberlin
- dlfkultur
- dlfnova
- dpa
- dpaAFX
- dpa_Kinder
- dpa_live
- dpa_unternehmen
- dpanord
- faznet
- focus_hamburg
- focusauto
- focusdigital
- focusfinanzen
- focusgesundheit
- focuskultur
- focusonline
- focuspanorama
- focuspolitik
- focusreise
- focussport
- focuswissen
- freie_presse
- handelsblatt
- hb_politik
- hessenschauDE
- heutejournal
- heuteshow
- hrinfo
- kn_online
- ksta_news
- lr_online
- mdr_th
- mdrde
- mediat_de
- mopo
- morgenmagazin
- morgenpost
- mozde
- mz_halle
- mz_landbote
- ndr2

- ndrmv
- neopresse
- neuepresse
- noz_de
- noz_el
- ntvNetzreporter
- ntvde
- ntvde_Politik
- ntvde_sport
- nwnews
- ots_bild
- ots_medien
- ots_people
- ots_politik
- ots_sport
- phoenix_de
- phoenix_kom
- presse_rbb
- presse_tp
- rbb24
- rbbFernsehen
- rbbabendschau
- rbbinforadio
- rheinpfalz
- rponline
- sportschau
- spox_news
- stern_sofa
- stern_sport
- sternde
- swr1bw
- swr3
- szaktuell
- szmagazin
- szonline
- tagesschau
- tagesthemen
- tazLesestoff
- tazTopStories
- taz_news
- tazamwe
- tazgezwitscher
- tvmainfranken
- wazwolfsburg
- wdr3
- wdr5
- welt
- westfalenblatt
- zeitonline
- zeitonline_dig
- zeitonline_kul
- zeitonline_pol
- zeitonline_wir
- zeitonline_wis
- zeitonlinesport
- zeitverlag

Appendix B

Redacted BASIL Example

ID: FOX00
Source: Fox News
Main event: Obama and Romney campaigns argue on Medicare
Main entities: Romney campaign, Obama campaign, Paul Ryan
Title: Ryan goes on offense over Medicare, accuses Obama of treating program like ‘piggy bank’
Url: `http://www.foxnews.com/politics/2012/08/14/[...]`
Word count: 1024
Stance

- **Main event**: Neutral
- **Romney campaign**: Neutral
- **Obama campaign**: Neutral
- **Paul Ryan**: Neutral

Sentence 1

- **Text**: Paul Ryan went on offense Tuesday in response to criticism over his Medicare plan, using an interview with Fox News coupled with a new TV ad to claim President Obama’s health care plan treats the treasured entitlement like a ‘piggy bank’ while the ‘Romney-Ryan’ plan preserves it.

- **Annotation 1**
  - **Start**: 150
  - **End**: 280
  - **Target**: Obama campaign
  - **Polarity**: Negative
  - **Aim**: Direct
  - **Indirect Target Name**: None
  - **Indirect Ally Opponent Sentiment**: None
  - **Bias**: Informational
  - **Quote**: No
  - **Speaker**: None
  - **Text**: President Obama' health care plan treats the treasured entitlement like a 'piggy bank', while the 'Romney-Ryan' plan preserves it"
- **Annotation 2**

  ...

Sentence 2

...

Appendix C

Training Parameters

Sentence-only baselines (Section 5.3):

BERT

- **Learning rate:** 2e-5
- **Batch size:** 16
- **Seeds:** 22, 20, 36, 64, 8
- **Epochs:** 10

RoBERTa

- **Learning rate:** 1e-5
- **Batch size:** 16
- **Seeds:** 49, 57, 33, 297, 181
- **Epochs:** 10

Context-Inclusive Models (ArtCIM & EvCIM, Section 5.4.1.2):

Stage 1: obtain sentence representations by taking the average of the last four layers of fine-tuned RoBERTa.

- **Model:** RoBERTa
- **Learning rate:** 1e-5
- **Batch size:** 16
- **Seed:** 49
- **Epochs:** 10

Stage 2: classify context representations obtained by processing sentence representations with a BiLSTM.

- **Model:** BiLSTM
- **Learning rate:** 1e-4
- **Batch size:** 32
- **BiLSTM layers:** 2
- **BiLSTM layer size:** 1000
- **Seeds:** 113, 114, 115, 116, 117
- **Epochs:** 75
- **Patience:** 5

Appendix D

BASIL IEF Targets by Commonness

IEF target commonness (Section 6.9) is measured in terms of the number of articles for which the entity appears in the corpus' main entity annotations. The top ten is marked in boldface.

- **Donald Trump** (85)
- **Barack Obama** (54)
- **Hillary Clinton** (24)
- **Republican Lawmakers** (20)
- **Democratic Lawmakers** (18)
- **Paul Ryan** (7)
- **House Democrats** (7)
- **Nancy Pelosi** (6)
- **Michael Flynn** (6)
- **Robert Mueller** (6)
- Vicente Fox (6)
- Benjamin Netanyahu (5)
- Adam Schiff (5)
- Senate Republicans (5)
- Bernie Sanders (4)
- U.S. Congress (3)
- Secure America Now (3)
- Roger Stone (3)
- Joe Lieberman (3)
- Ted Cruz (3)
- Trey Gowdy (3)
- Neil Gorsuch (3)
- Joe Biden (3)
- Mitt Romney (3)
- social media companies (3)
- Elizabeth Lauten (3)
- Sasha and Malia Obama (3)
- Rick Santorum (3)
- BuzzFeed (3)

- Michael Cohen (3)
- Catholics (3)
- Felipe Calder (3)
- Ruth Bader Ginsburg (3)
- John Boehner (3)
- Harry Reid (3)
- House Republicans (3)
- William Barr (3)
- Peter Strzok (3)
- GOP committee members (3)
- Michael Bloomberg (3)
- James Mattis (3)
- Rick Perry (3)
- Romney campaign (2)
- Roy Moore (2)
- National Enquirer (2)
- Senate Democrats (2)
- Kathy Griffin (2)
- Rashida Tlaib (2)
- AARP (2)
- social security (2)
- Republicans (2)
- State Department officials (2)
- Obama administration (2)
- Steve Scalise (2)
- Marco Rubio (2)
- Obama campaign (1)
- Tea Party (1)
- Enrique Pena Nieto (1)
- Michael Steele (1)
- Tim Ryan (1)
- senate Democrats (1)
- Cesar Sayoc (1)
- Law enforcement/Federal authorities (1)
- Stephen Fincher (1)
- Chuck Schumer (1)
- National Security Agency programs (1)
- Benghazi attacks (1)
- Democrats (1)
- Democrats presidential candidates (1)
- Victoria Nuland (1)
- Michael Morell (1)
- Obama’s administration (1)
- State Department (1)
- Susan Rice (1)
- lawmakers (1)

- White House officials and lawmakers (1)
- Jeb Bush (1)
- George W. Bush (1)
- Kevin McCarthy (1)

List of Figures

2.1 Illustration of *k*-means clustering. Initalization: data points are projected in a vector space and cluster centers are initialized to random points in the space. Iterations 1-3: data points are iteratively assigned to the center closest to them and centers are then re-computed to be the mean of their data points until the algorithm converges. 24

2.2 Diagram illustrating a Recurrent Neural Network in two ways. On the left, the RNN is depicted in its folded state. On the right, it is unfolded through time to reveal as many states as items in the input sequence X. At each time step *t* the RNN takes as input a fixed-size vector $\mathbf{x}_t$ and a hidden state $\mathbf{h}_t - 1$, and outputs a vector $\mathbf{o}_t$ which represents $\mathbf{x}_t$ and a hidden state $\mathbf{h}_t$ which represents the sequence from $\mathbf{x}_1$ until $\mathbf{x}_t$. The final hidden state, $h_{t=T}$ is returned as $\mathbf{o}_T$ to represent $\mathbf{x}_T$ or as z to represent the entire sequence $\mathbf{x}_T$, depending on the task the system is used for. 30

2.3 Diagram illustrating a BiLSTM for sequence classification: the BiLSTM encoder makes a forward and a backward pass over the input sequence. The resulting hidden states $\overrightarrow{\mathbf{h}_T}$ and $\overleftarrow{\mathbf{h}_T}$ are concatenated to obtain a encoded representation **z** and transformed by an MLP into a label **y**. 31

2.4 The Transformer architecture as depicted in Vaswani et al. (2017) 33

2.5 An illustration of a confusion matrix that represents model performance at a toxic language classification task. It distinguishes between true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). . . 34

5.1 Sentence-only baseline loss values on training data per epoch for different amounts of training data. 80

5.2 Sentence-only baseline F1-scores on validation data per epoch for different amounts of training data. 81

5.3 Diagram illustrating Sequential Sentence Classification as proposed by Cohan et al. (2019): the separator tokens are used as contextualised sentence representations by the classifier layer to generate labels for each sentence in the sequence. 83

5.4 Diagram illustrating Windowed Sequential Sentence Classification: sentence sequences are padded with the last sentence from the previous sequence and the first sentence of the subsequent sequence to prevent loss of context around sequence edges. 84

5.5 Diagram of the Article Context-Inclusive Model. In Stage 1, sentence embeddings are obtained by encoding padded word sequences using a pre-trained language model. In Stage 2, the sequence of sentence representations from the same news article as the target sentence is encoded by a BiLSTM, and concatenated with the target sentence representation. The result of the concatenation is classified with a linear classifier to obtain a sentence-level prediction. 86

5.6 Diagram of the Event-Context-Inclusive Model. In Stage 1, sentence embeddings are obtained by encoding padded word sequences using a pre-trained language model. In Stage 2, sequences of sentence representations from three articles on the same event are encoded by an equal number of BiLSTMs (only one shown in diagram). 87

List of Tables

4.1 Inter-annotator agreement for the on-target/off-target task (Krippendorff alpha) between the three FE workers and between the two experts who adjudicated tweets where FE worker judgment was not unanimous. 53

4.2 Adjudication for the on-target/off-target task of ETTC tweets where FE worker judgment was not unanimous. #Adj provides the number of adjudicated tweets. Diff. Expert 1 and 2 give the percentage of tweets where the expert label differed from the FE majority vote. 53

4.3 Number of tweets per politician per party in the GTTC. 55

4.4 Name and position on the political spectrum of the German political parties represented in the GTTC. 56

4.5 Possible responses to the stance annotation prompt. 58

4.6 Number of tweets, number of workers and stance annotator agreement (Krippendorff's alpha) for the ETTC after removing those annotators that were judged to be least reliable by MACE. 60

4.7 Examples of tweets with disagreeing annotations. Possible labels were: negative (-1), neither positive nor negative (0), positive (1) and cannot read / does not apply (x). In the first example, it is difficult to assess the stance of the tweet without knowing the content of the important message. In the second example, the tweet can be taken as an expression of sympathy, or as an ironic statement. 60

4.8 ETTC stance label distribution. 61

4.9 ETTC naming form distribution. 61

4.10 GTTC stance label distribution. 61

4.11 GTTC naming form distribution. 61

4.12 Possible examples of status indication and solidarity indication through naming in ETTC tweets, and an example of an alternative function. 62

4.13 Average stance and in parentheses the absolute number of tweets without or with a title in their naming form. 63

4.14 Average stance and in parentheses the absolute number of tweets by country subset and naming form from least to most formal. 64

4.15 Average stance and in parentheses the absolute number of tweets by politician (with party subtotals) and naming form from least to most formal. 65

4.16 Number and average stance of GTTC tweets that do or do not contain *Frau* (tweets with a female target) and *Herr* (tweets with a male target). In rows marker with "*", the difference is statistically significant ($p < 0.01$). 66

4.17 Number and average stance of GTTC tweets that do or do not contain *Dr.* (all tweets), *Bundeskanzlerin* or *Kanzlerin* (B/K) (tweets with target *Angela Merkel*) and *Professor* (P) (tweets with target *Jörg Meuthen*). 66

4.18 Examples of tweets with the title combination *Frau Dr.* and either positive or negative stance. 67

4.19 Number of GTTC tweets with likely left-leaning (LLL) or likely right-leaning (LRL) users and pro-left (PL), anti-right (AR), pro-right (PR) or anti-left (AL) stance. 68

4.20 Average naming formality in GTTC tweets with likely left-leaning (LLL) or likely right-leaning (LRL) users and pro-left (PL), anti-right (AR), pro-right (PR) or anti-left (AL) stance, and the naming formality gap as an absolute number and a percentage of negative tweets. 69

4.21 GTTC Addendum tweets by user leaning (left-leaning (LL) or right-leaning (RL)) and stance. 71

4.22 Average naming formality in the GTTC Addendum across left-leaning (LL) or right-leaning (RL) users and negative or positive stance , and the naming formality gap as an absolute number and a percentage of negative tweets. . 71

4.23 Quantities of tweets with and without title for mentions of female and male targets. 72

5.1 Example instances from a story in the BASIL corpus of informational bias (Fan et al., 2019) with their informational bias label (inf), news source (src) and BASIL ID. 76

5.2 Polarity labels of BASIL IEF instances. 78

5.3 Aim labels of BASIL IEF instances. 78

5.4 Quote labels of BASIL IEF instances. 78

5.5 Baseline performance without context for token and sentence classification. Sentences are divided into training and non-training sets by sentence or by story. We report standard deviations across 5 seeds for our models. Fan et al. (2019) report a minimum standard deviation of 3.36 and maximum of 12.44 for theirs. 80

5.6 Results of integrating direct textual context context with a Sequential Sentence Classifier without a window (SSC) or with a window (WinSSC) and a maximum sequence length of either 5 or 10. 85

5.7 Results of modeling article context with ArtCIM compared to a sentence-only baseline. 86

5.8 Results of incorporating event context with a Context-Inclusive Model compared to a RoBERTa sentence-only baseline. A dagger indicates that the improvement over the baseline is significant ($p < 0.05$). 88

5.9 Results of incorporating event context with a Context-Inclusive Model compared to a BERT sentence-only baseline. 88

5.10 Results of enhancing a Context-Inclusive Model with or without news source as an added feature. 89

5.11 Datasets used for extended pre-training: generic news domain data (DAPT), unlabeled BASIL data (TAPT), a combination of the two (DAPT-TAPT), and a selection of documents (where each document is a sentence) from the same source and time frame and with high similarity to the BASIL corpus (CurTAPT). 91

5.12 Results on IEF classification of BASIL data after adapting RoBERTa embeddings to generic news domain data (DAPT), unlabeled BASIL data (TAPT), a combination of the two (DAPT-TAPT), and to a selection of unlabeled sentences from the same source and time frame and with high similarity to the BASIL corpus (CurTAPT). 92

6.1 Example instances from three news stories in the BASIL corpus, along with their news source and labels for IEF (contains IEF: 1, does not contain IEF: 0), target (common: 1, uncommon: 0), polarity (positive: 1, negative: 0) and direction (direct: 1, indirect: 0). Note that Examples 2.3 and 3.2 contain sentiment-bearing language but are nevertheless considered instances of IEF because the opinion is contained within a direct quote. 95

6.2 RoBERTa and EvCIM performance (F1) and EvCIM's relative improvement over the baseline by sentence length (quartiles). The first column shows the number of tokens per sentence in each quartile (Q4 to Q1), the second the number of sentences in each quartile and the third the percentage of biased sentences per quartile. The final column shows the relative improvement of EvCIM over the baseline. A dagger indicates that the improvement is significant ($p < 0.05$). 97
6.3 Performance (F1) and EvCIM's relative improvement over the baseline on items that do or do not contain at least one strongly subjective token from the MPQA Subjectivity Lexicon. 98
6.4 Performance (recall) by commonness of the target entity. 99
6.5 Performance (F1) by whether instances' article features at least one common entity. 99
6.6 Performance (recall) by polarity. 100
6.7 Performance (recall) by direction. 100
6.8 Performance (recall) by whether the IEF span exists inside or outside of a quote. 101
6.9 Number of right-leaning, centrist and left-leaning articles from each news publisher. 101
6.10 Performance (F1) by the news outlet instances are from. 102
6.11 RoBERTa and EvCIM performance (F1) and EvCIM's relative improvement over the baseline by the stance of the instances' articles. 102

List of Terms and Abbreviations

agenda setting Theory which posits that the media influences the public by selecting which topics and which aspects of topics are to receive the most attention. 12, 14, 15

ANNs Artificial Neural Networks 28

BiLSTMs Bi-directional LSTMs 30, 85, 87

entity framing The framing of entities as opposed to topics. 4

FFNNs Feed-Forward Neural Networks 29

framing The process whereby the selection and presentation of aspects of a topic promote a particular interpretation of that topic. 1

Framing Theory The theoretical study of framing. 11

FT Framing Theory 11, 18

IEF Informational Entity Framing 2–9, 13, 77–79, 122

Informational Entity Framing The selection or omission of objectively verifiable aspects of an entity to promote a particular viewpoint towards that entity. 2

LDA Latent Dirichlet Allocation 25

LSTMs Long Short-Term Memory 30, 33, 34

Machine Learning A branch of computer science that develops algorithms which can learn from data in ways that seek to imitate human pattern recognition capabilities. 21

ML Machine Learning 21, 22

MLP Multi-layer Perceptron 31

Natural Language Processing The study of how computers can process, analyse and label large amounts of text and speech. 1

NB Naive Bayes 27

NLP Natural Language Processing . . . . v, 1–3, 7, 8, 11, 13, 21, 28, 31

NNs Neural Networks 28, 29

RNNs Recurrent Neural Networks 29, 30

SVM Support Vector Machine 27

References

Aarøe, L. and Petersen, M. B. (2020). Cognitive biases and communication strength in social networks: the case of episodic frames. *British Journal of Political Science*, 50(4):1561–1581.

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., and Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624.

Ali, W. and Hassoun, M. (2019). Artificial intelligence and automated journalism: Contemporary challenges and new opportunities. *International journal of media, journalism and mass communications*, 5(1):40–49.

Allerton, D. J. (1996). Proper names and definite descriptions with the same reference: A pragmatic choice for language users. *Journal of Pragmatics*, 25(5):621–633.

An, S.-K. and Gower, K. K. (2009). How do the news media frame crises? a content analysis of crisis news coverage. *Public relations review*, 35(2):107–112.

Antoine, J.-Y., Villaneau, J., and Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

Ba, J., Kiros, J., and Hinton, G. E. (2016). Layer normalization. *ArXiv*, abs/1607.06450.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

Bateson, G. (1955). A theory of play and fantasy. *Psychiatric Research Reports*.

Bateson, G. (1955/1973). A theory of play and fantasy. american psychiatric association psychiatric research reports 2, 1955. repr. *G. Bateson, Steps to an ecology of mind. London: Paladin*, pages 150–166.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., and Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Baumgartner, F. R., De Boef, S. L., and Boydstun, A. E. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

Berinsky, A. J. and Kinder, D. R. (2006). Making sense of issues through media frames: Understanding the kosovo crisis. *The Journal of Politics*, 68(3):640–656.

Besch, W. (1998). *Duzen, Siezen, Titulieren: zur Anrede im Deutschen heute und gestern*, volume 4009. Vandenhoeck & Ruprecht.

Bibas, S. (2004). Plea bargaining outside the shadow of trial. *Harvard Law Review*, pages 2463–2547.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Blommaert, J. et al. (2005). *Discourse: A critical introduction*. Cambridge University Press.

Boydstun, A. E., Card, D., Gross, J., Resnick, P., and Smith, N. A. (2014). Tracking the development of media frames within and across policy issues.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Brewer, P. R. (2003). Values, political knowledge, and public opinion about gay rights: A framing-based account. *Public Opinion Quarterly*, 67(2):173–201.

Brewer, P. R., Graf, J., and Willnat, L. (2003). Priming or framing: Media influence on attitudes toward foreign countries. *Gazette (Leiden, Netherlands)*, 65(6):493–508.

Broder, J. (2007). Familiar fallback for officials: ‘mistakes were made’. *New York Times*. URL: `https://www.nytimes.com/2007/03/14/washington/14mistakes.html`, Accessed 8 April 2021.

Brown, R. and Ford, M. (1961). Address in American English. *Journal of abnormal and social psychology*, 62(2):375–385.

Brown, R. and Gilman, A. (1960). The pronouns of power and solidarity. In Sebeok, T. A., editor, *Style in Language*, pages 253–276. MIT Press, Cambridge, MA.

Budak, C., Goel, S., and Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Buechel, S. and Hahn, U. (2017). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 578–585.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., and De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3):190–206.

Cabot, P.-L. H., Dankers, V., Abadi, D., Fischer, A., and Shutova, E. (2020). The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Findings of the association for computational linguistics: emnlp 2020*, pages 4479–4488.

Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 438–444.

Card, D., Gross, J., Boydstun, A., and Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420.

Carragee, K. M. and Roefs, W. (2004). The neglect of power in recent framing research. *Journal of communication*, 54(2):214–233.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, W.-F., Wachsmuth, H., Al Khatib, K., and Stein, B. (2018). Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88.

Choi, E., Tan, C., Lee, L., Danescu-Niculescu-Mizil, C., and Spindel, J. (2012). Hedge detection as a lens on framing in the gmo debates: A position paper. *arXiv preprint arXiv:1206.1066*.

Chong, D. and Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Clawson, R. A., Clawson, R., and Waltenburg, E. (2008). *Legacy and Legitimacy: Black Americans and the Supreme Court*. Temple University Press.

Clyne, M., Kretzenbacher, H.-L., Norrby, C., and Schüpbach, D. (2006). Perceptions of variation and change in German and Swedish address 1. *Journal of sociolinguistics*, 10(3):287–319.

Cohan, A., Beltagy, I., King, D., Dalvi, B., and Weld, D. S. (2019). Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

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

Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5640–5650.

Dahl, T. (2015). Contested science in the media: Linguistic traces of news writers' framing activity. *Written Communication*, 32(1):39–65.

Dang-Xuan, L., Stieglitz, S., Wladarsch, J., and Neuberger, C. (2013). An investigation of influentials and the role of sentiment in political communication on Twitter during election periods. *Information, Communication & Society*, 16(5):795–825.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

De Vreese, C. (2004). The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment. *Mass Communication & Society*, 7(2):191–214.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: empirical evidence. *Annu. Rev. Econ.*, 2(1):643–669.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., and Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Dernoncourt, F. and Lee, J. Y. (2017). Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

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

Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahendiran, A., Mille, S., Srivastava, A., Tan, S., et al. (2021). Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

Dickey, E. (1997). Forms of address and terms of reference. *Journal of linguistics*, 33(2):255–274.

Dimitrova, D. V., Kaid, L. L., Williams, A. P., and Trammell, K. D. (2005). War on the web: The immediate news framing of gulf war ii. *Harvard International Journal of Press/Politics*, 10(1):22–44.

Do, H. H., Prasad, P., Maag, A., and Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299.

Dowty, D. (1991). Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Druckman, J. N., Jacobs, L. R., and Ostermeier, E. (2004). Candidate strategies to prime issues and image. *The Journal of Politics*, 66(4):1180–1202.

Druckman, J. N. and Nelson, K. R. (2003). Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science*, 47(4):729–745.

Edwards, G. C. and Wood, B. D. (1999). Who influences whom? the president, congress, and the media. *American Political Science Review*, 93(2):327–344.

Edy, J. A. and Meirick, P. C. (2007). Wanted, dead or alive: Media frames, frame adoption, and support for the war in afghanistan. *Journal of communication*, 57(1):119–141.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Entman, R. M. et al. (2004). *Projections of power: Framing news, public opinion, and US foreign policy*. University of Chicago Press.

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. *Directions in sociolinguistics*, pages 213–250.

Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. *PloS one*, 9(2):e87908.

Fan, L., White, M., Sharma, E., Su, R., Choubey, P. K., Huang, R., and Wang, L. (2019). In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.

Field, A., Bhat, G., and Tsvetkov, Y. (2019). Contextual affective analysis: A case study of people portrayals in online# metoo stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 158–169.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580.

Fillmore, C. J. et al. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.

Fillmore, C. J. et al. (2006). Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

Fridkin, K. L. and Kenney, P. J. (2005). Campaign frames: Can candidates influence media coverage. *Framing American Politics*, pages 54–75.

Fulgoni, D., Carpenter, J., Ungar, L., and Preotiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3730–3736.

Galassi, A., Lippi, M., and Torroni, P. (2020). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gitlin, T. (2003). *The whole world is watching: Mass media in the making and unmaking of the new left*. University of California Press.

Glück, H. and Sauer, W. W. (1997). Gegenwartsdeutsch. 2., überarb. und erw. *Auf. Stuttgart/Weimar: Metzler (Sammlung Metzler. 252)*.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Greene, S. and Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511.

Greene, S. C. (2007). *Spin: lexical semantics, transitivity, and the identification of implicit sentiment*. PhD thesis.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Groseclose, T. and Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Günther, E. and Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International Journal of Communication*, 11:21.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Haidt, J. and Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Hamborg, F., Zhukova, A., and Gipp, B. (2019a). Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 196–205. IEEE.

Hamborg, F., Zhukova, A., and Gipp, B. (2019b). Illegal aliens or undocumented immigrants? towards the automated identification of bias by word choice and labeling. In *International Conference on Information*, pages 179–187. Springer.

Harden, B. (2006). On puget sound, it's orca vs. inc. *The Washington Post*, page A3.

Hartmann, M., Jansen, T., Augenstein, I., and Søgaard, A. (2019). Issue framing in online discussion fora. *arXiv preprint arXiv:1904.03969*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hickey, R. (2003). The German address system: Binary and scalar at once. *Pragmatics and Beyond*, pages 401–425.

Hiebert, R. E. (2003). Public relations and propaganda in framing the iraq war: A preliminary review. *Public Relations Review*, 29(3):243–255.

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

Hopper, P. J. and Thompson, S. A. (1980). Transitivity in grammar and discourse. *language*, pages 251–299.

Horne, B. D., Khedr, S., and Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth International AAAI Conference on Web and Social Media*.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Igartua, J. J., Cheng, L., and Muňiz, C. (2005). Framing latin america in the spanish press: A cooled down friendship between two fraternal lands. *Communications*, 30(3):359–372.

Iyengar, S., Kinder, D. R., et al. (1987). News that matters: Agenda-setting and priming in a television age. *News that Matters: Agenda-Setting and Priming in a Television Age*.

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122.

Jacoby, W. G. (2000). Issue framing and public opinion on government spending. *American Journal of Political Science*, pages 750–767.

Ji, Y. and Smith, N. (2017). Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829*.

Jiang, Y., Petrak, J., Song, X., Bontcheva, K., and Maynard, D. (2019). Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844.

Jin, D. and Szolovits, P. (2018). Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161*.

Johnson, K. and Goldwasser, D. (2018). Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730.

Johnson, K., Jin, D., and Goldwasser, D. (2017). Modeling of political discourse framing on twitter. In *Eleventh International AAAI Conference on Web and Social Media*.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Joseph, K., Friedland, L., Hobbs, W., Lazer, D., and Tsur, O. (2017). Constance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124.

Junk, W. M. and Rasmussen, A. (2019). Framing by the flock: Collective issue definition and advocacy success. *Comparative political studies*, 52(4):483–513.

Kaiser, J. and Kleinen-von Königslöw, K. (2019). Partisan journalism and the issue framing of the euro crisis: Comparing political parallelism of german and spanish online news. *Journalism*, 20(2):331–348.

Kako, E. (2006). Thematic role properties of subjects and objects. *Cognition*, 101(1):1–42.

Kang, S., Shim, K., and Kim, J. (2019). Social media posts on samsung galaxy note 7 explosion: A comparative analysis of crisis framing and sentiments in three nations. *Journal of International Crisis and Risk Communication Research*, 2(2):4.

Kato, Y., Kurohashi, S., Inui, K., Malouf, R., and Mullen, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research*.

Kause, A., Townsend, T., and Gaissmaier, W. (2019). Framing climate uncertainty: Frame choices reveal and influence climate change beliefs. *Weather, Climate, and Society*, 11(1):199–215.

Kellstedt, P. M. (2003). *The mass media and the dynamics of American racial attitudes*. Cambridge University Press.

Khanehzar, S., Turpin, A., and Mikolajczak, G. (2019). Modeling political framing across policy issues and contexts. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66.

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Kinder, D. R., Sanders, L. M., and Sanders, L. M. (1996). *Divided by color: Racial politics and democratic ideals*. University of Chicago Press.

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

Kulkarni, V., Ye, J., Skiena, S., and Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. *arXiv preprint arXiv:1809.03485*.

Lawrence, R. G. (2004). Framing obesity: The evolution of news discourse on a public health issue. *Harvard International Journal of Press/Politics*, 9(3):56–75.

Lecheler, S. and de Vreese, C. H. (2013). What a difference a day makes? the effects of repetitive and competitive news framing over time. *Communication Research*, 40(2):147–175.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Leech, G. et al. (1992). 100 million words of english: the british national corpus (bnc). *Language research*, 28(1):1–13.

Levy, K. E. and Franklin, M. (2014). Driving regulation: Using topic models to examine political contention in the us trucking industry. *Social Science Computer Review*, 32(2):182–194.

Lin, C., He, Y., and Everson, R. (2011). Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1153–1161.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. G. (2006). Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., and Lei, K. (2014). Sarcasm detection in social media based on imbalanced classification. In *International Conference on Web-Age Information Management*, pages 459–471. Springer.

Liu, S., Guo, L., Mays, K., Betke, M., and Wijaya, D. T. (2019a). Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

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

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Luo, Y., Card, D., and Jurafsky, D. (2020). Desmog: Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118.

Matthes, J. (2009). What's in a frame? a content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, 86(2):349–367.

McCombs, M. and Ghanem, S. I. (2001). The convergence of agenda setting and framing. *Framing public life: Perspectives on media and our understanding of the social world*, pages 67–81.

McCombs, M. and Reynolds, A. (2002). News influence on our pictures of the world. In *Media effects*, pages 11–28. Routledge.

McCombs, M. and Reynolds, A. (2009). How the news shapes our civic agenda. In *Media effects*, pages 17–32. Routledge.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

McCombs, M. E. and Shaw, D. L. (1993). The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of communication*, 43(2):58–67.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Mendelsohn, J., Budak, C., and Jurgens, D. (2021). Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*.

Menini, S., Nanni, F., Ponzetto, S. P., and Tonelli, S. (2017). Topic-based agreement and disagreement in us electoral manifestos. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2938–2944.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5528–5531. IEEE.

Miller, J. M. and Krosnick, J. A. (1996). News media impact on the ingredients of presidential evaluations: A program of research on the priming hypothesis. *Political persuasion and attitude change*, pages 79–100.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Minsky, M. (1974). A framework for representing knowledge.

Minsky, M. and Papert, S. (1969). Perceptrons.

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

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Moloney, K. (2006). *Rethinking public relations: PR propaganda and democracy*. Routledge.

Morris, A., Schammel, M., and Vissens, A. (2021). Talking sense: using machine learning to understand quotes. *The Guardian*. URL: `https://www.theguardian.com/info/2021/nov/25/talking-sense-using-machine-learning-to-understand-quotes`, Accessed 18 October 2022.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Naderi, N. and Hirst, G. (2017). Classifying frames at the sentence level in news articles. *Policy*, 9:4–233.

Nelson, T. E., Oxley, Z. M., and Clawson, R. A. (1997). Toward a psychology of framing effects. *Political behavior*, 19(3):221–246.

Neuman, W. R., Just, M. R., and Crigler, A. N. (1992). *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Nguyen, V. A. (2015). *Guided Probabilistic Topic Models for Agenda-Setting and Framing*. PhD thesis.

Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2013). Lexical and hierarchical topic regression. *Proceedings of Advances in Neural Information Processing Systems*.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., and Miler, K. (2015). Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448.

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, pages 798–808.

Nie, Y., Zhou, X., and Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Nisbet, M. C., Brossard, D., and Kroepsch, A. (2003). Framing science: The stem cell controversy in an age of press/politics. *Harvard International Journal of Press/Politics*, 8(2):36–70.

Noakes, J. A. and Wilkins, K. G. (2002). Shifting frames of the palestinian movement in us news. *Media, Culture & Society*, 24(5):649–671.

O'Connor, A. M. (1989). Effects of framing and level of probability on patients' preferences for cancer chemotherapy. *Journal of clinical epidemiology*, 42(2):119–126.

Opperhuizen, A. E., Schouten, K., and Klijn, E. H. (2019). Framing a conflict! how media report on earthquake risks caused by gas drilling: A longitudinal analysis using machine learning techniques of media reporting on gas drilling from 1990 to 2015. *Journalism Studies*, 20(5):714–734.

Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.

Palacio-Niño, J.-O. and Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Papalampidi, P., Keller, F., and Lapata, M. (2019). Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perloff, R. M. (2021). *The dynamics of political communication: Media and politics in a digital age*. Routledge.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Price, V., Tewksbury, D., and Powers, E. (1997). Switching trains of thought: The impact of news frames on readers' cognitive responses. *Communication research*, 24(5):481–506.

Rashkin, H., Bell, E., Choi, Y., and Volkova, S. (2017). Multilingual connotation frames: a case study on social media for targeted sentiment analysis and forecast. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–464.

Rashkin, H., Singh, S., and Choi, Y. (2015). Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.

Rasinski, K. A. (1989). The effect of question wording on public support for government spending. *Public Opinion Quarterly*, 53(3):388–394.

Raue, P.-J. (2012). Warum ein Dr. seinen Titel verliert. *Thüringer Allgemeine*. URL: `www.thueringer-allgemeine.de/leserinhalte/warum-ein-dr-seinen-titel-verliert-id218649719.html`, Accessed 21 November 2019.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1650–1659.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rhee, J. W. (1997). Strategy and issue frames in election campaign coverage: A social cognitive account of framing effects. *Journal of Communication*, 47(3):26–48.

Richardson, J. (2006). *Analysing newspapers: An approach from critical discourse analysis*. Palgrave.

Riker, W. H. R., Riker, W. H., Riker, W. H., and Mueller, J. P. (1996). *The strategy of rhetoric: Campaigning for the American Constitution*. Yale University Press.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. In *Lrec*, volume 2, pages 827–832. Las Palmas.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Roy, S. and Goldwasser, D. (2020). Weakly supervised learning of nuanced frames for analyzing polarization in news media. *arXiv preprint arXiv:2009.09609*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., and Choi, Y. (2017). Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Schank, R. C. and Abelson, R. P. (1977). Scripts, plans, goals, and understanding: an inquiry into human knowledge structures.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, 49(1):103–122.

Scheufele, D. A. (2000). Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society*, 3(2-3):297–316.

Scheufele, D. A. and Iyengar, S. (2012). The state of framing research: A call for new directions. *The Oxford Handbook of Political Communication Theories. New York: Oxford UniversityPress*, pages 1–26.

Scheufele, D. A. and Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20.

Schneider, A. and Ingram, H. (1993). Social construction of target populations: Implications for politics and policy. *American political science review*, 87(2):334–347.

Semetko, H. A. and Valkenburg, P. M. (2000). Framing european politics: A content analysis of press and television news. *Journal of communication*, 50(2):93–109.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 8.

Siegal, A. M. and Connolly, W. G. (1999). *The New York Times manual of style and usage*. Three Rivers Press (CA).

Slothuus, R. and De Vreese, C. H. (2010). Political parties, motivated reasoning, and issue framing effects. *The Journal of Politics*, 72(3):630–645.

Sniderman, P. M. and Theriault, S. M. (2004). The structure of political argument and the logic of issue framing. *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*, pages 133–65.

Soroka, S. and McAdams, S. (2015). News, politics, and negativity. *Political Communication*, 32(1):1–22.

Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B., and Donnay, K. (2021). Mbic–a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.11910*.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Taddy, M. (2013). Measuring political sentiment on Twitter: Factor optimal design for multinomial inverse regression. *Technometrics*, 55(4):415–425.

Takiff, H. A., Sanchez, D. T., and Stewart, T. L. (2001). What's in a name? the status implications of students' terms of address for male and female professors. *Psychology of Women Quarterly*, 25(2):134–144.

Tankard, J., Hendrickson, L., Silberman, J., Bliss, K., and Ghanem, S. (1991). Media frames: Approaches to conceptualization and measurement (paper presented to the annual meeting of the association for education in journalism and mass communication). *Boston, Massachusetts*.

Tourni, I., Guo, L., Daryanto, T. H., Zhafransyah, F., Halim, E. E., Jalal, M., Chen, B., Lai, S., Hu, H., Betke, M., et al. (2021). Detecting frames in news headlines and lead images in us gun violence coverage. In *Findings of the Association for Computational Linguistics: 2021 Conference on Empirical Methods in Natural Language Processing. November 2021, pages 4037-4050, Punta Cana, Dominican Republic.*

Tsur, O., Calacci, D., and Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1629–1638. ACL.

Tuchman, G. (1980). *Making News*. Free Press.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 178–185.

Túñez-López, J.-M., Fieiras-Ceide, C., and Vaz-Álvarez, M. (2021). Impact of artificial intelligence on journalism: Transformations in the company, products, contents and professional profile. *Communication & Society*, 34(1):177–193.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Uscinski, J. E. and Goren, L. J. (2011). What's in a name? coverage of senator hillary clinton during the 2008 democratic primary. *Political Research Quarterly*, 64(4):884–896.

van Atteveldt, W. and Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, 12(2-3):81–92.

van den Berg, E., Korfhage, K., Ruppenhofer, J., Wiegand, M., and Markert, K. (2019). Not my President: How names and titles frame political figures. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 1–6.

van den Berg, E., Korfhage, K., Ruppenhofer, J., Wiegand, M., and Markert, K. (2020). Doctor who? Framing through names and titles in german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4924–4932, Marseille, France. European Language Resources Association.

van den Berg, E. and Markert, K. (2020). Context in information bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Van Dijk, T. A. (1998). Discourse and ideology. *Discourse and Society*, 9:307–308.

Van Gorp, B. (2007). The constructionist approach to framing: Bringing culture back in. *Journal of communication*, 57(1):60–78.

Van Gorp, B. (2010). Strategies to take subjectivity out of framing analysis. In *Doing news framing analysis*, pages 100–125. Routledge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *ICWSM*, pages 454–463.

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24.

Weaver, D. H. (2007). Thoughts on agenda setting, framing, and priming. *Journal of communication*, 57(1):142–147.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Wicke, P. and Bolognesi, M. M. (2020). Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *arXiv preprint arXiv:2004.06986*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308.

Wiegreffe, S. and Marasović, A. (2021). Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Yano, T., Resnik, P., and Smith, N. A. (2010). Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158. Association for Computational Linguistics.

Zaller, J. R. et al. (1992). *The nature and origins of mass opinion*. Cambridge university press.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Ziems, C. and Yang, D. (2021). To protect and to serve? analyzing entity-centric framing of police violence. *arXiv preprint arXiv:2109.05325*.

Zoizner, A. (2018). The consequences of strategic news coverage for democracy: A meta-analysis. *Communication Research*, page 0093650218808691.