DOES MEDIA BIAS AFFECT TWITTER OUTREACH?

A MULTI-FEATURE ANALYSIS OF NEWS ARTICLES

ELISABETH RICHTER

Universität Konstanz

Master Thesis submitted in partial fulfilment of the requirements
for the degree of Social and Economic Data Science
Faculty of Sciences
Department of Economics
Universität Konstanz

Evaluated by Jun.-Prof. Juhi Kulshrestha Ph.D.
Evaluated by Prof. Dr. Bela Gipp

14.03.2022

Elisabeth Richter
Brüelstraße 11
78462, Konstanz

REFEREES:
1. Referee: Jun.-Prof. Juhi Kulshrestha Ph.D.
2. Referee: Prof. Dr. Bela Gipp

SUPERVISOR:
Timo Spinde

LOCATION:
Konstanz

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Timo Spinde for his support and guidance over the past few months and through whom the project has only become possible.

Furthermore, I would like to thank Dr. Dipl.-Psych. Helge Giese, who helped me with his statistical skills. Thank you for the always prompt and detailed feedback to my questions.

I would also like to express my gratitude and admiration to Smilla Hinterreiter, who impressed me with her graphic talent and helped me by creating a beautiful visualization of the media bias framework.

The media bias framework was developed in close collaboration with Timo Spinde and Felix Blochwitz. I would like to thank both of them for this successful teamwork.

Lastly, I also want to thank: Markus for encouraging and motivating me throughout the last months; Gunda for sharing all known Overleaf tips and tricks (and for providing me with delicious Swiss chocolate); and all of my friends and family who have supported me and did some proof-reading.

ABSTRACT

Nowadays, many information exchanges take place online, and social media platforms like Facebook or Twitter are popular mediums for all kinds of information gathering. Unsurprisingly, news outlets increasingly rely on online media channels to disseminate their news stories. But basically, anyone can write stories and share them with a large potential readership. This uncontrolled flow of information can be problematic, leading to increased amounts of misleading or inaccurate information circulating.

Such inaccurate or false news is only one of many manifestations of a phenomenon called media bias. Media bias refers to all kinds of biased, i.e., non-neutral, news, including, for example, one-sided or inaccurate news coverage or reporting in a way to harm or favorite a person or object of interest intentionally. There is a lot of research on media bias, which shows that the manifestations of media bias tend to be multi-layered and complex. However, the existing research provides valuable insights without considering the bigger picture. Often, it is not clear how different media bias types are related to each other, where similarities exist, and where they differ significantly, making evaluating current discoveries difficult.

In general, this thesis has the overall goal to better understand the structure of media bias. Therefore, the first contribution is to define a theoretical framework for media bias that potentially guides future research more concretely. Based on a detailed literature review, different media bias categories are developed, and different bias types are assigned to these categories. In addition, this framework allows better embedding of media bias within the context of other related concepts.

Furthermore, the attempt is made to examine whether the article's comments can be an indicator of its bias. As no dataset suitable for this purpose yet exists, large amounts of user-generated Twitter data as well as article- and outlet-specific metrics of ad fontes media are collected. By that, a new multi-layered media bias dataset is created, valuable for sophisticated media bias research.

Based on this data, a multi-feature analysis is conducted to identify comment characteristics and whether they are indicators of an article's bias. The comments are examined regarding their level of hate and their sentiment polarity. By that, media bias is set into context with other related concepts. The results show that the more hateful the comments on an article are, the higher its level of bias. Furthermore, these results are underpinned by the finding that the news outlet's individual stance reinforces this hate-bias relationship.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

Ever since the Internet became available to the broad public, new information and communication technologies have been established. Suddenly, the world became more connected, and news articles were shared faster and with broader reach, which eventually reshaped the entire media environment [22]. News consumption shifts more and more towards the online world. According to the *Digital News Report 2021* [102], the use of traditional print newspapers experienced a sharp decline, similar to the development of the previous years. In contrast, the use of social media platforms for news is very high, especially among younger users [102]. However, the use of Facebook, Twitter, or other platforms is problematic, as "[t]he share of social media engagements from unreliable news sites doubled from 2019 to 2020" [21, p. 1]. Often, misleading information is shared online, with no one feeling responsible for regulating this flow of news [21]. Examples for fake news exist many, including the false claims during and after the 2020 US election, misinformation about climate change, or, just recently, the spread of conspiracy theories about COVID-19 [21]. Therefore, it is no surprise the trust in media is rather low. According to a study conducted by the Reuters Institute, only 44% of the respondents stated "[...] they trust most news most of the time" [102, p. 9].

The occurrence of misinformation, fake news, or inaccurate reporting can be summarized under the term media bias. The phenomenon of media bias is defined as "[...] slanted news coverage or internal bias, reflected in news articles" [127, p. 2]. A biased news article usually "[...] leans towards or against a certain person or opinion by making one-sided, misleading or unfair judgements" [84, p. 1268]. Biased news coverage happens "[...] when journalists report about an event in a prejudiced manner or with a slanted viewpoint" [86, p. 1]. Media bias has severe consequences for our society, as biased news eventually manipulate voting behavior [35, 39, 111] or influence political decisions [13, 65, 81]. Some scholars even fear that a biased news coverage endangers democratic values [65, 81, 83]. Hence it is not surprising that a large part of (online) news consumers fear being exposed to fake news [102]. The problem of false, misleading, or inaccurate news reporting is not an issue of the "modern world". Research on media bias goes back at least until the 1950s [138], but the topic gains more relevance with each year.

As news stories are increasingly read online, consequently, news consumers are confronted with a more diverse media environment. The World Wide Web offers almost unlimited amounts of news stories from all sorts of news outlets [110]. Here, the decision of which news outlet to trust is mainly the responsibility of the individual news consumer [65]. But how does one know if the trusted news outlet really can be trusted? Or how does one detect the trustworthy articles among a flood of news stories? In order to answer these questions and to properly judge whether an article is biased or not, the news consumers must genuinely understand the news, which in turn requires unbiased news reporting [65]. This is why media bias research is so important: to break this vicious circle.

1.2 CONTRIBUTIONS AND RESEARCH OBJECTIVES

With regard to existing literature, a lot of research has already been conducted. Many researchers made an attempt to define the concept of media bias and outline its characteristics. Reviewing this work underlines that media bias is a complex phenomenon that manifests in many different ways. For example, a common definition of media bias divides the concept into the following three subcategories: gatekeeping bias, coverage bias, and statement bias [30, 67]. However, this definition is not universal, and other types of media bias are mentioned, for example, ideology bias and spin bias [98]. Often, researchers provide valuable insights but do not consider the bigger picture. The missing of a universal theoretical framework for media bias is a significant downside of current media bias research. This work aims to fill this gap by presenting a framework for media bias that allows for a more straightforward classification of different media bias types. By defining several media bias subcategories and assigning different bias types to these categories, this thesis aims to reduce the concept's complexity and improve future research.

While studying the existing literature on media bias, some concepts have been detected that are often mentioned together with media bias. Two of these concepts are hate speech and sentiment analysis. Hate speech is defined as any kind of "[...] language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" [33, p. 512]. Much research on hate speech has been conducted, most of them concentrating on hateful language on social media platforms or within online communities.

Sentiment analysis, on the other hand, is a text processing technique from the area of natural language processing and enables the detection of emotions, moods, or opinions in human language [16, 45]. Sentiment analysis is valuable for many tasks, for example, to predict movie reviews or stock market behavior [45]. Similar to hate speech, sentiment analysis is also often applied to short textual data like posts or comments on social media platforms [41]. In some cases, sentiment analysis can even be used as a technique to detect hate speech [2, 112].

Often, the underlying goal of hate speech detection and sentiment analysis tasks is to understand the dynamics of online communities and to observe the impacts hateful language might have [41, 119]. However, searching through the literature exposes that in most cases, the underlying data are social media data, and only very little attention has been paid to observing user comments on news articles [143]. This thesis pursues the goal of deepening the research on understanding media bias. The overall aim is to connect the three concepts media bias detection, sentiment analysis, and hate speech detection. As of now, to the best of my knowledge, no such research has been conducted yet. With this project, it is aspired to fill the gap by observing the following research questions:

RQ1: Are user comments on a news article an indicator of the article's bias?

RQ2: Is the new outlets' stance an indicator of the article's bias?

In order to answer these questions, data on statement-level (i.e., user comments) and data on article-level are required, with the statements being explicitly linked to the respective article. An extensive review of existing datasets for media bias detection, sentiment analysis, and hate speech detection shows that no suitable dataset exists for this purpose. Thus, to be able to deduct the analysis, appropriate data is required to be collected in the first step.

Hence, the three main contributions of this work are as follows:

C1: The development of a universal theoretical framework for media bias detection - the media bias framework.

C2: The construction of a first-of-its-kind dataset useful for a combined study of media bias, sentiment analysis, and hate speech.

C3: Conducting an analytical study by observing comment characteristics on news articles in order to detect indicators of the article's bias.

1.3 STRUCTURE OF THE THESIS

This work is structured as the following. First, Chapter 2 provides an extensive overview of the existing literature on media bias. The literature review includes the comparison of several common definitions of media bias and outlines the impacts of biased news coverage. A major part of Chapter 2 consists of presenting the media bias framework. Additionally, an overview of state-of-the-art approaches on automated media bias detection, sentiment analysis, and automated hate speech detection is provided, as well as a listing of all relevant datasets applicable to the respective concepts. Lastly, the gaps of the current media bias research are pointed out and based on these, the research objectives of this thesis are derived.

In Chapter 3 an in-depth explanation of the methodological procedure is provided. First, the data collection process is described, consisting of the collection of outlet-related and article-related metrics as well as the collection of user comments. Second, the approaches for analyzing the characteristics of the collected comments are presented and discussed. Lastly, the statistical models are described, which best suits the underlying data structure to obtain the most valuable insights.

Chapter 4 reports the results of the three steps described in Chapter 3. This includes a detailed description of the obtained dataset as well as the results of the statistical analysis. In addition, the findings that are derived from the results are presented.

In Chapter 5, the limitations of this work are thoroughly discussed, and an outlook for future research is provided.

Lastly, Chapter 6 concludes the thesis.

2 RELATED WORK AND THEORETICAL EMBEDDING

This chapter presents an overview of the existing literature on media bias. In Section 2.1 the term media bias is defined, and the impacts media bias has, are explained. The following Section 2.2, presents the current state-of-the-art research of media bias. Based on the existing literature, a theoretical framework is defined. This framework divides media bias into several subcategories and assigns different media bias types to these categories. It also includes additional concepts closely related to media bias research. Section 2.3 presents the current state-of-the-art approaches for automated detection of media bias, as well as for two additional techniques, sentiment analysis, and hate speech detection. In Section 2.4, an overview of existing datasets for media bias detection, sentiment analysis, and hate speech detection is provided. Section 2.5 points out existing gaps in the current media bias research and states the contributions of this work. Lastly, in Section 2.6 the research objectives of this thesis are outlined, and the theoretical embedding is explained.

2.1 UNDERSTANDING MEDIA BIAS

When reviewing existing literature on media bias, it quickly becomes apparent that the whole idea of media bias is complex. Among scholars, many definitions of media bias circulate, each differing based on the particular context of their work. Research on media bias goes back at least until the 1950s [138]. Consequently, many attempts have been made to tackle bias in the news. Thus, many researchers define either the whole concept of media bias or only one of its many manifestations. Some of these definitions of media bias have consolidated over the years.

Firstly, a widespread definition of media bias is the division into three major categories: gatekeeping bias, coverage bias, and statement bias [30, 65]. According to this definition, media bias can express itself in one or more of these categories. Gatekeeping bias refers to the process where "[...] writers and editors select from a body of potential stories those that will be presented to the public [...]" [30, p. 135]. The term stems from the idea that media representatives act as "gate keepers", making the final decision of what news stories are reported [30, 113, 138]. Obviously, these decisions are also logistical, as simply not all events are of significant importance. However, due to personal preferences, these decisions can be influenced and thus result in biased reporting [30, 98, 113, 138]. Second, coverage bias describes the situation where two or more sides of an issue receive imbalanced amounts of attention, for example, pro-life vs. pro-choice statements [30]. In general, an article's attention is measured as the length of the article, how much space it gets in a newspaper (e.g., printed on the front page), or how often the topic is reported on [30, 113]. Lastly, statement bias refers to when "[...] members of the media [...] interject their own opinions into the text [...]" [30, p. 136]. This division is popular among scholars, although sometimes, the terms used to describe each bias type might differ. For example, gatekeeping bias is sometimes also called selection bias or agenda bias, coverage bias is sometimes called visibility bias, and statement bias is sometimes referred to as tonality bias [43]. Another term for biased reporting is editorial slant, which functions as an umbrella term for gatekeeping, coverage, and selection bias [39].

Another common definition of media bias defines only two types of biases: ideology bias and spin bias [98]. Ideology bias describes the situation where the editor's or journalist's personal preferences influences what events and how these events are reported [98]. This is often in line with the news outlet's stance, "[f]or example, left wing newspapers may simply prefer to report news one way" [98, p. 2]. Spin bias, on the other hand, describes the situation where an event is being reported on, but potentially relevant information is omitted [92, 98]. This is a common practice among newspapers and stems "[...] from a newspaper's attempt to tell a simple and memorable story" [98, p. 2], in order to attract as many potential readers as possible [5].

Interestingly, the definition of ideology bias is quite similar to the ideas behind gatekeeping and statement bias combined. This example convincingly shows one of the major issues that comes with defining the media bias concept, namely that the same ideas are named differently and that sometimes the ideas of several media bias types can overlap.

In some cases, media bias is also referred to as a kind of lexical or linguistic bias [15]. These approaches have their roots mainly in linguistics and are based on the idea of the linguistic category model introduced by Semin & Fiedler [118] in 1988. The main idea is that the language we use differs in its level of abstraction, depending on whether a particular behavior was anticipated or not. The decision to formulate things using abstract language happens mostly subconsciously [15] and, compared to the other bias types introduced above, independent of the context of the event. However, the use of abstract language might reinforce existing stereotypes [15, 90].

Considering the above-presented definitions of media bias, it becomes clear that they tend to refer to rather specific types of media bias. A universal, generally valid definition of media bias does not really exist, but occasionally, researchers have defined media bias more generally. For example, Spinde, et al. [127] define media bias as "[...] slanted news coverage or internal bias reflected in news articles" [127, p. 2]. Lazaridou, et al. [84] on the other hand, define media bias as news reporting that "[...] leans towards or against a certain person or opinion by making one-sided misleading or unfair judgements" [84, p. 1268]. Lee, et al. [86] define media biased as reporting "[...] in a prejudiced manner or with slanted viewpoint" [86, p. 1].

Moreover, media bias does not only manifests via text but also the way the news is presented to the reader contributes to the biased news coverage [66, 93]. For example, portraying a person with happy facial expressions is perceived as favorable, whereas portraying a person with angry facial expressions is rather perceived as dominant [104].

The impacts of the existence of media bias are far-reaching. Almost all individuals are affected by media bias, and its consequences [65]. In the very early stages of media bias research, it has already been observed that the mass media has political impacts and, in turn, influences the voting behavior [111]. These findings are underpinned by additional work, confirming the influencing effects of news coverage on political tendencies, which in turn impacts voters' behaviors [35, 39] and voter turnout [53, 132].

Complementary work provides even deeper insights by showing that media bias might even influence the decision-making process of individuals and thereby leads to individuals' making electoral mistakes [13, 81]. This stems from the individual's tendency to adopt biased views when being exposed to biased news coverage [65, 83]. A survey conducted by Kull, et al. [83] shows that misinformation about the Iraq war led to individuals' supporting the war. In addition, news consumers who perceive media as biased tend to have a lower level of trust in the government [69].

Furthermore, biased news coverage and public discussion about (controversial) topics might increase the incidence of extremism. This is due to a phenomenon called "group polarization", which states that discussions with like-minded others lead to reinforcing existing beliefs, which in turn leads to more extreme opinions [128]. As a consequence, "[...] the polarization of public opinion [...] complicates agreements on contentious topics" [65, p. 393]. In addition, biased reporting might reinforce stereotypical mindsets [90].

All the impacts mentioned above are further reinforced when individuals are trapped in so-called "echo chambers". This is the case when news consumers are mainly "[...] surrounded by news and opinions close to their own" [65, p. 392]. Especially in the context of social media platforms and the existing recommender system, users tend to read only those news stories that confirm their existing beliefs [65, 99, 100]. Hence, it is no surprise that some scholars even fear that media bias threatens the stability of democracies [65, 81, 83].

A lot of research on media bias and its manifestations exists. What has been demonstrated above is that researchers often focus on different types of media bias instead of the entire concept. Some research concentrates on media bias expressing in form of linguistic cues [15, 59, 73, 90, 107, 108]. Another group of scholars rather focuses on context-dependent forms of bias, either on reporting level [19, 30, 50, 113], or on text level [5, 19, 30, 48, 73, 92, 98]. Lastly, a large group of existing work observes what factors influence the personal perception of news. Many of these researchers refer to the hostile media effect, which describes the phenomenon that each group perceives media as biased towards their own point of views [63, 131], and define influencing factors reinforcing this effect. For example, the political ideology [32, 47, 57, 87] or the level of partisanship are strong indicators for the hostile media effect [47, 57, 109], but other factors are objects of research as well [6–9, 11, 23, 29, 42, 60–62, 70, 85, 93, 104, 116, 122, 123, 130, 142, 144].

In some of these works, theories and definitions overlap, and it is not always possible to draw a clear distinction between particular bias types. A further difficulty is that some scholars refer to the same concept but name it differently, which makes a summary of the state-of-the-art research even more challenging. Nevertheless, due to the complexity of the concept of media bias, it is particularly necessary to be able to clearly distinguish its characteristics. This thesis aims to fill this gap by specifying a media bias framework that enables a clear classification of different media bias types. This media bias framework is elucidated in more detail in the following Section 2.2.

2.2 THE MEDIA BIAS FRAMEWORK

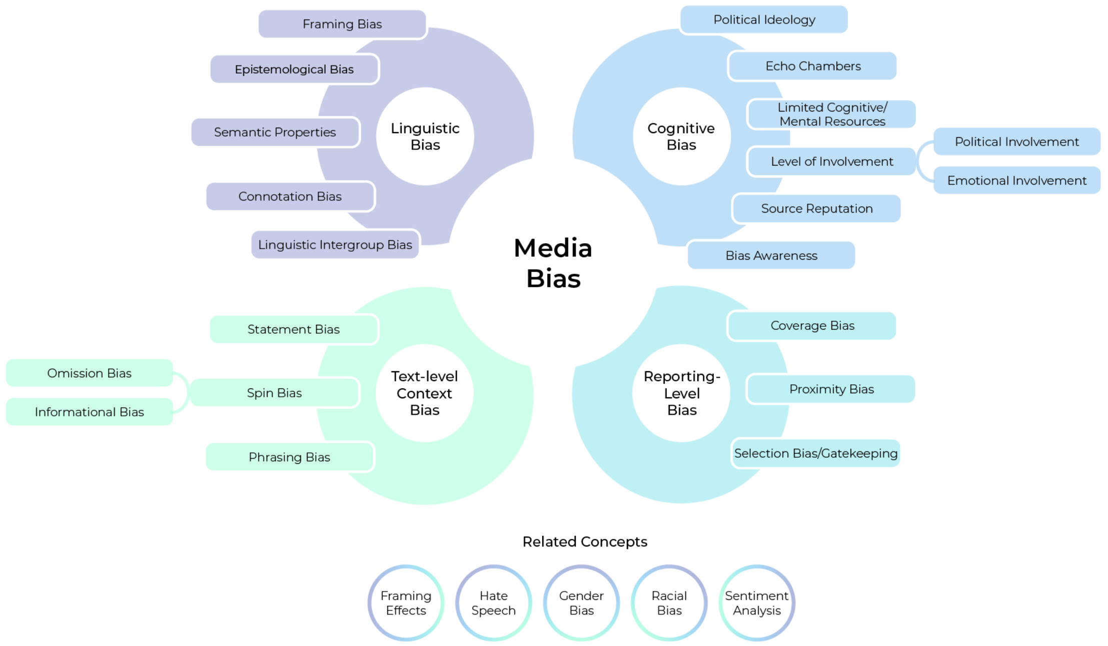
In order to gather all relevant work on media bias, an extensive keyword-based literature research has been conducted, using Google Scholar[1] and KonSearch[2]. First, all work related to media bias has been searched (keyword: "media bias") but eventually narrowed down by using more specific search terms (e.g., "media bias in newspapers", "media bias coverage", or "media bias perception"). This has already returned the majority of literature that has been considered in this thesis. However, more research has been found by observing cited work of papers that have been considered to be especially relevant. The search has not been limited to a specific time period, as it is important for the development of a comprehensive understanding of media bias to also learn about the beginnings of media bias research. A total of about 200 scholarly papers and related works have been found, of which about one-third are considered to be particularly relevant. Most of them will be referred to in the following Section.

In Figure 2.1 the theoretical framework of media bias is visualized. The concept is divided into four major subcategories: linguistic bias, text-level context bias, reporting-level context bias, and cognitive bias. In addition, the fifth category related concepts refer to five concepts that have been encountered while investigating the media bias literature. These concepts are no bias types *per se*. Hence, they cannot be exclusively assigned to one of the four bias categories. However, for the sake of completeness, it is necessary to mention and explain these terms. In the following, the definitions for all four subcategories and their corresponding bias types, as well as for the related

[1] https://scholar.google.com/

[2] https://konstanz.summon.serialssolutions.com/; the literature search engine owned by the University of Konstanz

Figure 2.1: The Media Bias Framework



Note: This Figure visualizes the theoretical framework for media bias. Each category embodies a subcategory of media bias and consists of different bias types. Concepts mentioned under "Related Concepts" are not bias types *per se*, but are closely related to media bias.

concepts, are provided.

2.2.1 *Linguistic Bias*

The first subcategory, linguistic bias, is defined as "[...] a systematic asymmetry in word choice that reflects the social-category cognitions that are applied to the described group or individual(s)" [15, p. 1]. In other words, biases of this type are triggered due to lexical features like word choice and sentence structure. Linguistic bias mainly focuses on the question of how things are said and is independent of the context. This kind of bias often reflects stereotypical mindsets [90] and is rather subconscious [15]. In the literature, it is sometimes also referred to as lexical bias [48]. Five bias types have been identified that belong to this category: linguistic intergroup bias, framing bias, epistemological bias, bias by semantic properties, and connotation bias.

Linguistic Intergroup Bias

The linguistic intergroup bias is a concept based on the linguistic category model (LCM) defined by Semin & Fiedler [118] in 1988. The main idea of the LCM is that words are categorized into one of four classes, depending on their level of abstraction. These classes are descriptive action words, interpretive action words, state verbs, and adjectives, where the first is the least abstract class, the latter the most abstract [38, 118]. Maass & Salvi [90] then defined the term linguistic intergroup bias, which is based on the idea that depending on the anticipated behavior of in-group and out-group members, more abstract or more concrete language is used. Maass & Salvi [90] illustrate this with the following example, considering the hypothetical scenario where "Person A is hitting Person B's arm with his fist" [90, p. 982]. Describing this scenario using the least abstract form of language, one could say, "A is punching B" [90, p. 982]. This entails no kind of valuation or implication and only describes what has happened. In contrast, using the most abstract form of language, one could say "A is aggressive" [90, p. 982]. This might or might not be true and cannot be judged from the pure fact that A hit B [90]. The use of such language is often subtle and reinforces stereotypes [15, 90].

Framing Bias

Framing bias is defined as the use of "[...] subjective words or phrases linked with a particular point of view" [108, p. 1650] and by that swaying the meaning of a statement [108]. Such subjective words are either one-sided terms or subjective intensifiers [108]. One-sided terms are words that "[...] reflect only one of the sides of a contentious issue" [108, p. 1653], for example "pro-life" vs. "anti-abortion" [108]. Both terms refer to the same movement, however, from completely opposite viewpoints. Subjective intensifiers are adjectives or adverbs that reinforce the meaning of a sentence, for example, "fantastic" vs. "accurate" [73, 108]. According to Recasens, et al. [108], the occurrence of framing bias is rather explicit.

Epistemological Bias

Epistemological bias describes the use of "[...] linguistic features that subtly [...] focus on the believability of a [statement]" [108, p. 1650]. For example, using the word "stated" conveys higher veracity than the word "claimed" [108]. Word classes associated with epistemological bias are factive verbs, entailments, assertive verbs, and hedges [108]. Factive verbs are verbs that indicate truthfulness [108]. For example, "He realized that..." indicates that the ensuing statement is true [108]. In contrast, "He thinks that..." does not imply the truthfulness of the following statement. Entailments are relations where one word implies the truth of another word [108]. These relations are directional, which means they only hold in one direction. For example, the word "murder" entails "kill" because murdering someone implies that this person has been killed, whereas the opposite direction is not true, as "killing" does not necessarily imply "murder" [108]. Another feature is assertive verbs which, as the name already indicates, asserts a statement [108]. This is closely related to factive verbs but focuses more on the neutrality of a statement. For example, verbs like "say" or "state" are usually perceived as neutral, whereas "claim" indicates doubt [108]. Lastly, hedges are certain words used to introduce vagueness to a statement [108]. For example, words like "may", "possibly", or "eventually" make a sentence less committed to the truth [108]. In contrast to framing bias, epistemological bias is rather subtle and implicit [108].

Bias by Semantic Properties

This type of bias, defined by Greene & Resnik [59], explains how so-called "semantic properties" trigger bias [59]. Similar to framing bias and epistemological bias, bias by semantic properties is also based on the idea that the way how something is framed changes its meaning. The difference, however, is that framing and epistemological bias refer to *what* words are used, whereas bias by semantic properties refers to *how* the sentence is structured. Greene & Resnik [59] explain this by means of the following example. The sentence "A soldier veered his jeep into a crowded market" [59, p. 503] puts the focus on the soldier as the executive character. By slightly reframing the sentence to "A soldier's jeep veered into a crowded market" [59, p. 503], the focus shifts from the solider to an unknown object. The actual statement of the sentence remains the same - a jeep crashed into a market - but without particularly blaming anyone for it [59].

Connotation Bias

Lastly, connotation bias refers to the idea that the use of connotations introduces bias to a statement [107]. To fully understand this type of bias, one must know the difference between the denotative meaning of a word and its connotative meaning. The denotation of a word is its literal meaning. The connotation, however, refers to a secondary meaning besides its literal meaning and is usually linked to certain feelings or emotions associated with a particular point of view [107]. For example, "undocumented worker" and "illegal alien" both have the same denotation, i.e., both words literally mean the same group of people [137]. However, the connotations, i.e., the secondary meanings, of the two words are certainly different [137].

2.2.2 *Text-level Context Bias*

Similar to linguistic bias, this subcategory also refers to the question of how something is said. The difference is, however, biases of this category consider the statement's context. Based on the context, certain words or statements included in the article alter the context and, by that, influence the reader's opinion [5, 19, 30, 48, 73, 92, 98]. Bias types belonging to this category are statement bias, phrasing bias, and spin bias, which in turn consists of omission bias and informational bias.

Statement Bias

Statement bias refers to when "[...] members of the media [...] interject their own opinions into the text[...] " [30, p. 136], which in turn leads to certain news being reported in a way that is more or less favorable towards a particular position [30]. These opinions can be very faint and are expressed "[...] by disproportionately criticizing one side[...] " [19, p. 250] rather than "[...] directly advocating for a preferred [side] [...]" [19, p. 250].

Phrasing Bias

This type of bias is characterized by the use of inflammatory words, i.e., language that is non-neutral [73]. The difficulty is that depending on the context, a word can change from being neutral to being inflammatory. Hence, the inter-dependencies between words and phrases must be considered, and whether the statement becomes more neutral when the word is exchanged [73]. For example, not every statement containing the word "murder" is biased (cf. "He was convicted of murder" [73, p. 4]). However, depending on the context, "murder" becomes an indicator for bias (cf. "An abortion is the murder of a human baby" [73] vs. "An abortion is the intentional ending of a pregnancy"[3]).

Spin Bias

Spin bias describes a form of bias introduced either by leaving out necessary information [92, 98], or by adding unnecessary information [5, 48]. This stems "[...] from [the] newspaper's attempt to tell a simple and memorable story" [98, p. 2], in order to attract as many potential readers as possible and "[...] to survive in the media market" [5, p. 531].

Spin bias can further be divided into omission bias and informational bias. Omission bias is defined as the act of omitting words from a sentence [92], which is also called simplifications [92, 98]. Informational bias is defined as the act of adding speculative, tangential, or irrelevant information to a news story [48], which is also referred to as exaggerations [5, 48].

[3] https://dictionary.cambridge.org/us/dictionary/english/abortion; accessed on 2022-02-25

2.2.3 *Reporting-level Context Bias*

The reporting-level context bias category assembles all types of bias that occur on the reporting level. Compared to the text-level context bias, which observes bias within an article, bias types of this category observe reasons that trigger unequal coverage of, or imbalanced attention for certain topics [19, 30, 50, 113]. Types of bias belonging to this category are selection bias, proximity bias, and coverage bias.

Selection Bias
Selection bias, also called gatekeeping bias, refers to the selection process where "[...] writers and editors select from a body of potential stories [...]" [30, p. 135]. Obviously, not all news events can be reported on due to the limited resources of the newspapers. However, these decisions are also prone to bias as personal preferences might influence this decision-making process [30, 98, 113, 138].

Coverage Bias
Coverage bias describes the situation where two or more sides of an issue receive imbalanced amounts of attention, for example, pro-life vs. pro-choice statements [30]. The level of attention can be measured either, for example, in absolute numbers (e.g., there are more articles discussing pro-life than pro-choice topics), or as how much space the topics get in a newspaper (e.g., printed on the front page), or as the length of the article (e.g., pro-life articles are longer and receive more in-depth coverage than pro-choice articles) [30, 113].

Proximity Bias
This type of bias, similar to coverage bias, also refers to certain events or topics receiving less attention than others. The difference, however, is that proximity bias focuses on cultural similarity and geographic proximity as decisive factors. Newspapers tend to report more frequently and more in-depth on events that happened nearby [113]. Additional evidence exists that the more culturally similar a country is, the more likely it is that events from that region or country will be reported, and the coverage will be more in-depth [50, 113]. On the contrary, the more culturally distinct a country is, the more likely it is that only big events are reported (e.g., natural disasters), or that only stereotypical news is reported (e.g., political instability) [50]. In his work, Galtung [50] also observes that countries usually report on the same group of foreign countries. For example, the U.S. tend to report more extensively on Latin America, whereas Great Britain rather reports on the British Commonwealth countries.

2.2.4 *Cognitive Bias*

Cognitive bias is defined as "[...] a systematic [...] deviation from rationality in judgment or decision-making" [18, p. 1]. In other words, cognitive bias refers to systematic errors in the decision-making that lead to irrational judgments [18]. Such misjudgments potentially influence how news consumers perceive the news and lead to different perceptions of media bias. A concept that is closely related to the perception of media bias is the hostile media effect. This term describes the phenomenon where members of opposing groups both rate the news article as biased against their own point of views [131]. Many research exists that observe the factors influencing the appearance of the hostile media effect and its strength. Consequently, these papers are highly media-focused. Apart from that, research on cognitive bias examines its causes on a more abstract level. From these two viewpoints, factors are derived that explain the cognitive bias in the context of media bias. These factors are political ideology, the existence of echo chambers, the level of involvement, which is further divided into political involvement and emotional involvement, the source reputation, the level of bias awareness, and limited cognitive or mental resources.

Political Ideology

The political ideology of individuals shapes the way how news is perceived [32]. The literature provides evidence that Republicans are more likely to distrust the media compared to Democrats and hence, are more likely to perceive media as biased [47, 57, 87].

Existence of Echo Chambers

The social environment of an individual influences the individual's decision-making [18]. With regard to media bias, researchers observe that discussions with like-minded others and the existence of echo chambers (i.e., being in a bubble "[...] where only certain ideas, information and beliefs are shared" [42, p. 729]) lead to a significantly greater perception of bias [42, 47]. This phenomenon has also been observed in the online environment, for example, by Houston, et al. [70], and by Lee [85], both finding that user comments influence how news articles are perceived.

Level of Involvement

From a broader perspective, one cause of cognitive bias is the individual's emotions and its level of affection, as they shape its decision-making [18]. Furthermore, existing literature provides evidence that the level of involvement has an influence on media bias perception. Hence, the following two factors have been derived from multiple existing research: the level of political involvement and the level of emotional involvement.

First, the level of political involvement refers to the level of partisanship, i.e., how strongly an individual is committed to a political party. Existing work shows that those with stronger party affiliation were more likely to perceive media as hostile towards their own point of view [47, 57]. In addition, evidence exists that higher group assimilation leads to higher partisanship which in turn leads to stronger hostile media effect [109].

Second, the level of emotional involvement refers to how strongly individuals identify themselves with a particular topic. For example, gun owners are more likely to perceive an article about a stronger gun control law as hostile towards their point of view because they are emotionally more involved than non-gun owners [144]. This personal consternation, also with the potential consequences of such a stronger gun control law, consequently leads to stronger emotions in general, which in turn influences the perception of bias [7, 144].

Source Reputation

Existing research shows that balanced articles are perceived as more or less hostile, solely depending on the media outlet [6, 8, 62, 130]. On the one hand, this applies to news sources that are incongruent with the individual's political orientation [142]. For example, a study shows that Muslims perceived articles in favor of Muslims when they were associated with a Muslim newspaper but biased against Muslims when it was associated with a Christian newspaper [6]. On the other hand, this effect also refers to the outlet's general reputation. For example, news stories published by Fox News are generally perceived to be biased towards Republican's point of view as Fox News is known to be in favor of conservatives the Republican party [11].

Level of Bias Awareness

The level of bias awareness refers to the individual's knowledge about bias. Spinde, et al. [122] observe that the way how bias is presented to the readers influences their bias awareness. This means presenting readers an article where biased sentences are marked improves the reader's awareness about that bias. In a complementary work, Spinde, et al. [123] show that annotators' training on media bias detection significantly increases annotation quality. Consequently, obtaining higher awareness of bias in the media, either through well-chosen visualizations of bias or through specific training, has an influencing effect on the perceptions of bias. Thus, the strength of the hostile media effect decreases, as skilled readers are less likely to falsely perceive news as biased.

Limited Cognitive or Mental Resources

Another cause for cognitive bias is the "[...] limited processing capacity of the human mind" [18, p. 2]. Sometimes, the human brain is not capable of processing all cognitive stimuli and experiences a mental overload. As a result, "[...] the mind uses [...] mental shortcuts [...]" [18, p. 4] to jump to a conclusion without making a rational decision [18]. With regard to the media environment, this is the case when news consumers are faced with an oversupply of news stories. This informational overload results in individuals reading only a small subset of newspapers [65] or dealing with only one side of an issue. Consequently, this might reinforce the perception of media bias, as the risk of sliding into an echo chamber or collecting only one-sided information increases.

2.2.5 *Related Concepts*

Concepts belonging to this category are no bias types *per se* but are often mentioned in the context of media bias. The definitions and how they relate to media bias are explained in the following.

Framing Effects

The concept of framing effects is based on the idea that media discourse is a dynamic process that is structured in frames [132], i.e., "[...] interpretive packages that give meaning to an issue" [51, p. 3]. As a consequence, such frames might lead "[...] to promote a particular interpretation" [46, p. 164] or to highlight certain aspects while others are overlooked [46, 51]. By that, frames and the resulting framing effects partly change the way how news consumers perceive news. Therefore, framing effects can be viewed as a generic term that comprises all the biases resulting from news frames.

Hate Speech

Hate speech is defined as any kind of "[...] language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" [33, p. 512]. Hateful language can manifest in different ways, for example, statements that mean to derogate, insult, or humiliate a person based on stereotypical characteristics like gender, nationality, sexual orientation, or religion [33, 97, 134]. Consequently, hateful language includes various types of biased language. For example, a statement containing gender bias can also be considered to be hateful language. Embedding hate speech within the media bias context, a logical implication is that any statement or article that contains hateful language is considered to be biased. In other words, hateful language is an indicator of bias. However, hate speech is no separate bias type because hateful language can manifest through different bias types, for example, through statement bias, phrasing bias, or connotation bias, to only name a few. Thus, it is most accurate to consider hate speech as a related concept. More precisely, to declare hate speech as a bias mechanism that introduces bias to articles or statements. The consequences of hate in media content are severe, as it might reinforce tension or hatred between groups or countries and thus encourage violent activities [120]. Existing work on hate speech is wide-ranging, but most research focuses on online content, e.g., user-comments of Facebook or Twitter [2, 33, 97, 112, 119, 134, 145]. Existing approaches for hate speech detection will be further elaborated in Section 2.3.3.

Gender Bias

Gender bias is defined as the "[...] dominance [...] of one gender over the other" [27, p. 495], which leads to "[...] the less dominant gender [being] underrepresented and stereotypes appear [...]" [27, p. 495]. As already mentioned in the above paragraph, gender bias can be a form of hate speech. However, in the context of media bias, gender bias is also expressed in different forms, apart from language cues. In the literature, sufficient evidence for the existence of gender bias in the media exists. For example, men, compared to women, are significantly more often the main character of a news story, leading to significantly lower coverage about female characters in general [3, 34]. Davis [34] also observes that news stories featuring female characters are usually "[...] shorter stories and those with smaller headlines [...]" [34, p. 458]. In addition, women are quoted significantly less often than men and are identified more often "[...] by personal information such as attire, physical description, [or] marital and parental status [...]" [34, p. 457].

At this point, it is important to highlight that Davis's work is from 1982, which is 40 years ago, and justifies the question of whether his findings are still valid today. With ongoing discussions about gender equality and women's empowerment, one might assume that news coverage about women changed and brought into line with that of men. However, more recent work shows there is still an unequal level of coverage of males and females in the news. In 2010, Ali, et al. [3], also find that men are significantly more often reported on than women, but they outline that the level of unequal coverage is different within different topics. For example, "[...] articles about sports or business are among the most gender-biased, while articles in entertainment are the least gender-biased" [3, p. 37].

More work on unequal coverage of women exists, for example, the one by Min & Feaster [95], who observe that the reporting on missing children is less for girls than boys. The examples stated are clearly some sort of coverage bias or selection bias. Nevertheless, gender bias does not only manifest via those two forms exclusively but can also be expressed in the form of bias types assigned to linguistic bias or text-level context bias [31]. To demonstrate this, consider Davis's [34] example that women are more often described via personal information. Adding, for example, the marital status to a news story only because the person in question is female and not because the information is relevant for understanding the story is a sort of spin bias. Another exemplification stems from the fact that the public generally expects nurses to be female and doctors to be male [15, 27]. These stereotypical expectancies may introduce bias to the news in the form of linguistic bias types. Lastly, unequal levels of coverage, as described by Ali, et al. [3], might be a form of selection bias or coverage bias. Consequently, gender bias should not be assigned to one bias category or bias type only. Similar to hate speech, gender bias should also be considered as bias mechanism that manifests through different bias types.

Racial Bias

Racial bias is defined as the systematic disproportionate under- or overrepresentation of minority groups in a specific context [95]. The idea behind racial bias and gender bias is relatively similar. Hence, the following two arguments that have been stated for gender bias also apply to racial bias. For one thing, this is that racial bias can be a form of hate speech, and second, that in the context of media bias, racial bias can take on different forms of bias. Also here, sufficient evidence exists in the literature for the existence of racial bias in the media environment. However, the separation of racial bias and gender bias is not always as clear. For example, Gershon, et al. [54] observe that "[...] minority congresswomen often receive more negative and less frequent media coverage than all other representatives" [54, p. 105]. Here, the authors combine racial and gender factors and come to the conclusion that the observed behavior only refers to minority congresswomen. Being a woman or being a minority alone does not have such an impact [54].

A similar research objective has been observed by Min & Feaster [95], who have discovered that the reporting on missing children is heavily biased based on their race and gender. Their work presents evidence that the news covered the missing of girls significantly less than the missing of boys as well as the missing of African American children less than the missing of non-minority children [95]. These findings contradict the findings of Gershon, et al. [54] and allow the implication that both factors, gender, and race, have impacts independent of each other.

In addition to that, other scholars discovered that minority groups were more often presented as criminals, i.e., in a bad light, whereas majority groups were rather presented as victims, i.e., in a good light [26, 37]. The reasoning why racial bias is considered as a related concept follows a similar pattern as for gender bias and hate speech. Consequently, also racial bias is considered as a bias mechanism that introduces bias through different bias types.

Sentiment Analysis

Sentiment analysis is a natural language task and refers to the process of analyzing text with respect to its emotional content, mood, or opinion [16, 45]. The sentiment of a statement contains great informational values, which are helpful to various amounts of problems [74]. For example, past research has applied sentiment analysis for all kinds of tasks, for example, to predict movie reviews or stock market behavior [45], or to observe how citizens of different countries reacted to the COVID-19 outbreak at the beginning of 2020 [41].

In the context of media bias, sentiment analysis can be used to detect bias in statements or articles. For example, Enevoldsen & Hansen [45] have studied if sentiment analysis can be used to detect political bias in Danish newspapers. The results show that the articles about the left-wing party contain more positive sentiment than the articles about the right-wing party [45]. A similar approach has been pursued by Hamborg & Donnay [64], who build a model to detect target-dependent sentiment (TSC) classification in newspaper articles. TSC is a sentiment analysis task that aims to find the polarity towards a target [64]. In addition to that, sentiment analysis can be also be helpful to detect hate speech, which, as derived above, is one form of biased language [2, 112]. For example, Rodríguez, et al. [112] used sentiment and emotion analysis to detect hate speech on Facebook. Hube & Fetahu [72] made the facile assumption that any statement that contains some kind of sentimental polarity is consequently not neutral, and therefore, the statement is considered to be biased.

The existing literature on sentiment analysis demonstrates that it can be a valuable task to observe bias in the media environment. Consequently, sentiment analysis can be considered as a detection mechanism for the identification of media bias. Further sentiment analysis approaches will be discussed in more detail in Section 2.3.2.

2.2.6 *Summary: Theoretical Framework*

In summary, what has been discussed in this Section is that media bias is a complex phenomenon that manifests through different bias types and utilizes several mechanisms. In order to provide a neat framework, the concept has been divided into four categories. First, linguistic bias consists of linguistic intergroup bias, framing bias, epistemological bias, bias by semantic properties, and connotation bias. Bias types of this category influence *how* a statement is framed independent of the sentence's context, typically through linguistic or grammatical cues. Second, bias types belonging to text-level context bias are statement bias, phrasing bias, and spin bias, which in turn splits into omission bias and informational bias. This category also affects *how* a statement is framed, but, in contrast to linguistic bias, here, the statement's context plays a role. The third category, reporting-level context bias, views media bias on a more abstract level. The bias types belonging to this category are selection bias, proximity bias, and coverage bias and influence *what* is reported on. Fourth, cognitive bias refers to how news consumers perceive news and what factors lead to a differing perception of media bias. These factors are the political ideology, the existence of echo chambers, the level of political or emotional involvement, the source reputation, the level of bias awareness, and the available cognitive or mental resources. Depending on how strong these factors are, individuals perceive news as more or less biased. Lastly, several related concepts have been aggregated into a fifth category called related concepts. These related concepts are no bias types *per se* and hence cannot be assigned to one media bias category exclusively. Concepts belonging to this category are framing effects, hate speech, gender bias, racial bias, and sentiment analysis.

2.3 EXISTING APPROACHES

The preceding Section 2.1 provides a detailed definition of the concept of media bias. The drawback of the current state-of-the-art research has been outlined, which is the missing of a universal theoretical framework for media bias research. To fill this gap, the media bias framework has been presented. In addition, media bias has been embedded within the context of other related concepts. Two of these concepts are considered to be particularly important: sentiment analysis and hate speech. As outlined above, hateful language is an indicator of bias. Hence, hate speech detection is relevant when studying media bias. In addition, sentiment analysis is a valuable technique to detect media bias [45, 64, 72] and hateful language [112]. As a logical implication, sentiment analysis and hate speech detection are relevant related concepts. In the following Section, a brief summary of the current state-of-the-art approaches for the three concepts is provided.

2.3.1 *Automated Media Bias Detection*

A glance at the literature reveals that many different approaches for the automatic detection of media bias exist. Roughly, existing work can be split into two areas: those that apply machine learning algorithms to detect bias and those that apply more advanced self-learning models.

Such machine learning approaches have, for example, been applied by Recasens, et al. [108], who trained a logistic regression model in order to detect bias-inducing words in a biased sentence. Similar to that, Hube & Fetahu [72] also trained a supervised model to detect bias on sentence level. A slightly more sophisticated approach has been proposed by Baumer, et al. [12], who adopt a variety of machine learning algorithms to detect biased language. For example, Stochastic Gradient Descent, Logistic Regression, or k-Nearest Neighbor has been adapted, with the result that Naive Bayes performs best [12]. Testing multiple algorithms and comparing their performance is a common procedure. Spinde, et al. [127] also have adopted various machine learning approaches to detect bias-inducing words on article level. For example, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), or Logistic Regression algorithms have been implemented with the result that XGBoost, a decision tree implementation, performs best [127]. Lastly, another machine learning approach has been proposed by Chen, et al. [24], who applied a Gaussian Mixture Model to improve bias classification.

In contrast to that, the use of self-learning language models has become increasingly popular. Applying deep learning approaches like Recurrent Neural Networks (cf. [25, 73]) or transformer models like BERT (cf. [48, 125]), ELECTRA (cf. [125]), or XLNet (cf. [125]) allow for a more advanced classification approach. These neural models enable sequential text processing, which allows capturing the contextual embedding of a word [73] and by that improves bias detection. In addition, the fine-tuning of existing language models with multi-task learning approaches leads to improved media bias detection [124]. Spinde, et al. [124] argue that multi-task learning is especially valuable in cases of scarce data, which applies to media bias detection as high-quality datasets are rare.

Lastly, alternative approaches exist, for example, by Hamborg, et al. [67] who have pursued an NLP-based approach, where word embeddings have been used to detect bias in news articles.

2.3.2 *Sentiment Analysis*

Studying the literature for sentiment analysis reveals that the number of existing approaches can be divided into three main areas. First, the most basic approaches for sentiment analysis are lexical approaches. The main idea is to create word lexicons that contain for each word its polarity (positive, negative) [16, 40]. In the literature, some of these dictionaries are referenced. First, there are dictionaries like LIWC[4], Harvard's General Inquirer (GI)[5], or Cambridge's Hu-Liu04[6] lexicon, which contain a binary label for each word's polarity (positive, negative) [74]. However, other dictionaries exist where words are linked to sentiment intensity, for example, SentiWordNet or SenticNet [74]. Given such lexicons, the overall polarity of a sentence can then be determined by adding up each word's polarity [16]. For example, Enevoldsen & Hansen [45] applied such a dictionary-based approach to detect the sentiment of Danish newspaper articles. The authors derived the mean sentiment score for each article based on the Danish sentiment dictionary AFINN [45, 103]. Just recently, Dubey [41] adapted a lexical approach to classify Twitter data on COVID-19 with regard to sentiment. The author used the R package "syuzhet"[7], which "[...] classifies the tweets on the basis of sentiments (positive and negative) and also categorizes them into 8 emotions [...]" [41, p. 3].

Another large area of sentiment analysis research focuses on the application of machine learning approaches to detect bias. Common algorithms like Naive Bayes, Support Vector Machines, Ensemble Learner, or Maximum Entropy have been used to detect a statement's sentiment (cf. [58, 68, 101]). The existing literature shows that the most frequently used algorithms for sentiment detection are Support Vector Machines and Naive Bayes [16]. Some scholars also propose combined approaches [40], for example, Joyce & Deng [80] using the OpinionFinder Lexicon [139] together with Naive Bayes algorithm to get the sentiment of tweets.

[4] https://www.liwc.app/

[5] http://www.mariapinto.es/ciberabstracts/Articulos/Inquirer.htm

[6] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

[7] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Similar to media bias detection, an increasing amount of researchers implement more advanced deep learning approaches to detect sentiment. For example, Wang, et al. [133] have proposed a CNN-LSTM model that is able to differentiate partial information from the text input. Other neural network-based approaches consider variations of Recurrent Neural Networks (cf. [1]) or applied combined attention-based approaches (cf. [10]). In order to overcome shortcomings of, for example, Recurrent Neural Networks, Transformer based architectures are proposed (cf. [52, 77]).

2.3.3 *Automated Hate Speech Detection*

Given the large amounts of hate speech research, existing work can be divided into three categories. The first category refers to rule-based approaches [4, 96]. The idea behind these methods is that the text is classified for hate based on a set of rules. These rules are manually created but enriched by word lists and linguistic cues, for example, grammar, morphology, or semantics [4, 96]. For instance, Gitari, et al. [56] build a rule-based classifier to detect hate speech, using three features: word polarity, a hate lexicon, and grammatical patterns [56]. Such rule-based approaches usually achieve high accuracy but are very time-intense [4]. Rodríguez, et al. [112] introduce a different approach whose goal is to detect hate speech on Facebook. To achieve this, the authors first deduce a sentiment analysis and then clustered the posts accordingly [112].

Another part of existing research applies machine learning algorithms to build a hate speech classifier [96]. Similar to media bias detection and sentiment analysis, standard algorithms are used, for example, Support Vector Machines (cf. [119, 134]) or Logistic Regression (cf. [33]). In addition, existing literature proposes that hate speech is a multi-class classification problem rather than binary class. For example, Davidson, et al. [33] point out that supervised machine learning methods are error-prone as they fail to detect the slight differences between hateful language and other kinds of offensive language.

The third category of automated hate speech detection refers to deep learning approaches [96]. The state-of-the-art approaches for hate speech detection are similar as already mentioned for media bias detection and sentiment analysis. For example, approaches exist where variations of Convolutional Neural Networks are implemented (cf. [145, 146]) or where multi-task learning approaches are combined with a BERT-based language model [97].

2.3.4 *Summary: State-of-the-Art Approaches*

What has been outlined in this Section 2.3 is that similar state-of-the-art approaches for all three concepts exist. This is no surprise as media bias detection, sentiment analysis, and hate speech detection are all three a subtask of text classification. From a broader perspective, supervised machine learning and deep neural models are promising approaches for the three concepts. In general, the implementation of transformer language models has a wide range of applications, where media bias detection, sentiment analysis, and hate speech detection are only a small variety. In most cases, these more sophisticated models outperform the supervised learning approaches.

2.4 EXISTING DATASETS

As demonstrated in Section 2.3 a large variety of different text classification approaches exists. These techniques usually require labeled data for the training or fine-tuning process of the particular algorithm. However, collecting sufficient amounts of data and getting high-quality annotations is difficult [91, 124]. In the following, a listing of existing datasets for media bias detection, sentiment analysis, and hate speech detection is provided.

2.4.1 *Media Bias Datasets*

For media bias research, most of the existing datasets provide annotations on sentence level [25, 49, 72, 73, 88, 108] with some having additional information on word-level [12, 125, 126], but a few exceptions provide labels on article-level [28, 48, 89]. A few of the articles are domain-specific, i.e., referring to only one news event or topic [12, 28, 49, 89], but most of the data refers to several topics. In the following, an overview is provided.

- ***NPOV Corpus - 1 [108].*** NPOV stands for "neutral point of view" and refers to Wikipedia's policy to ensure that all articles are written from a fair and non-biased point of view[8]. Recasens, et al. [108] created a dataset by extracting all articles that had been in Wikipedia's category of NPOV disputes[9] and then further split the articles into sentences. Each article consists of a set of revisions, i.e., different versions of the article. The authors then extracted all words that had been changed from one version to the other, called edits. Hence, each edit contains of a before word and an after word.

- ***NPOV Corpus - 2 [73].*** Hube & Fetahu [73] also created a dataset by collecting all Wikipedia articles that had been in Wikipedia's category of NPOV. The authors then asked crowdsource workers to annotate each statement with regard to the presence of bias. The final dataset contains statements labeled as biased or neutral.

[8] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

[9] https://en.wikipedia.org/wiki/Category:All_Wikipedia_neutral_point_of_view_disputes

- ***Wikipedia dataset [72].*** Hube & Fetahu [72] created a dataset by collecting the latest Wikipedia articles with each statement being labeled as biased or unbiased.
- ***Framing dataset [12].*** Baumer, et al. [12] published a domain-specific dataset by collecting political news articles. The final dataset consists of a number of articles where each word has been annotated for framing.
- ***Single-event articles [89].*** This dataset created by Lim, et al [89] contains news articles to one single event, "Black men arrested in Starbucks", that has been annotated with the labels not biased, slightly biased, fairly biased, and strongly biased.
- ***Ukraine crisis dataset - 1 [28].*** Cremisini, et al. [28] present a dataset that contains news articles that are related to the Ukraine crisis. Each article was manually rated as either pro-Russian, pro-Western, or neutral.
- ***Ukraine crisis dataset - 2 [49].*** Färber, et al. [49] extended the previously mentioned dataset by Cremisini, et al. [28] by annotating it on sentence-level with regard to bias dimensions. These dimensions are Hidden Assumptions and Premises, Subjectivity Framing, and Overall Bias.
- ***MBIC [126].*** MBIC stands for Media Bias Including Characteristics and is a dataset compiled by Spinde, et al. [126]. This dataset consists of sentences that had been annotated for bias with labels on sentence-level and word-level. On sentence-level, annotators had been asked to label the sentence as biased or non-biased and if the sentence expresses an opinion. On word-level, annotators had to state which the bias-inducing words are.
- ***BABE [125].*** BABE stands for Bias Annotations By Experts and is a dataset that was built on top of MBIC [125, 126]. It contains sentences that have been labeled on sentence-level and word-level by expert annotators. The annotator-schema was the same as for MBIC [125].
- ***BASIL [48].*** BASIL stands for Bias Annotation Spans on the Informational Level and is a dataset created by Fan, et al. [48]. The authors collected sets of articles where each set refers to a similar news event. The articles have been annotated on the article level with labels on the overall polarity and on sentence-level with labels on bias type, bias target, bias polarity, bias aim, and whether the biased statement is a quote.
- ***LimA [88].*** This dataset, conceived by Lim, et al. [88] consists of sentences referring to four different events. The sentences had been annotated by crowdsource workers with labels on the sentence's bias.

- ***ChenA [25].*** Chen, et al. [25] created this dataset by first collecting articles from *allsides.com*[10], containing labels on topic and political bias. Then the authors added a fairness label provided by ad fontes media. Lastly, each sentence had been labeled either as political bias, unfairness, or non-objectivity, where the labels were in turn derived from the ad fontes media labels.

2.4.2 *Sentiment Analysis Datasets*

The majority of existing sentiment datasets are compiled of Twitter data [58, 114, 117, 129], only the dataset created by Hamborg & Donnay refers to newspaper articles [64]. However, the annotation schema varies between the different datasets. Some contain sentiment labels on tweet-level [58, 117, 129], whereas others sentiment information on target-level [64, 114]. Also, the terms used to label sentiment and the number of sentiment classes are inconsistent. The most general case is labels for positive, negative, and neutral [58, 129], but sometimes sentiment classes are added, for example, a label for mixed sentiment or other [114].

- ***Stanford Twitter Sentiment (STS)[11] [58].*** This sentiment dataset has been compiled by Go, et al. [58] and contains tweets labeled as positive, negative, or neutral. It consists of a training set and a test set, with the test set being manually labeled.

- ***STS-Gold [114].*** This corpus has been constructed based on the STS dataset, but annotations are on tweet and target level. The number of labels have been expanded by adding the two additional labels mixed, and other [114].

- ***Sentiment Strength Twitter dataset (SS-Twitter)[12] [129].*** This dataset has been constructed by Thewall, et al. [129] and contains tweets that have been annotated for their sentiment strength. This means, labels for negative sentiment are on a scale from -1 (not negative) to -5 (extremely negative), and on a scale from 1 to 5 for positive sentiment respectively [129].

- ***SemEval datasets[13] [117]*** SemEval is a once-a-year international workshop on semantic evaluation. As part of the event, the respective datasets are provided for download. For example, the SemEval2014 dataset[14] contains Twitter data which have been labeled on tweet-level as positive, negative, or neutral [78]. Datasets from other years exist as well [114].

[10] https://www.allsides.com/unbiased-balanced-news
[11] http://help.sentiment140.com/home
[12] http://sentistrength.wlv.ac.uk/documentation/
[13] https://semeval.github.io/
[14] https://alt.qcri.org/semeval2014/task9/index.php?idDdata-and-tools

- ***NewsMTSC*[15] *[64].*** This dataset created by Hamborg & Donnay [64] contains sentences that were annotated with labels on the sentence's target and whether the polarity towards the target is positive, negative, or neutral.

2.4.3 *Hate Speech Datasets*

The vast majority of hate speech datasets consist of social media data like Twitter [33, 135, 136] or Stormfront [55], a white supremacist forum. Some of the data have binary labels [55], although the majority of annotations consist of three or four labels [33, 36, 117, 135, 136]. To ensure transparency, the names for the datasets in the following list have been adopted from MacAvaney, et al. [91].

- ***HatebaseTwitter*[16] *[33].*** This dataset has been created by Davidson, et al. [33] and consists of Twitter data. The tweets have been labeled as hate speech, offensive language, or neither.

- ***WaseemA*[17] *[136].*** This dataset also contains of Twitter data and was annotated with labels for racist, sexist, or neither.

- ***WaseemB*[18] *[135].*** With this dataset, Waseem [135] extended the previously created dataset by Waseem & Hovy [136]. The annotation schema is similar to WaseemA, except that one more label has been added. Hence, this dataset categorizes Twitter data into racist, sexist, neither, or both.

- ***Stormfront*[19] *[55].*** This dataset contains binary annotated data from Stormfront. The labels are either hate or no hate.

- ***HatEval*[20] *[117].*** This dataset is from the SemEval competition 2019 (Task 5) and is annotated on hate and aggression, and additionally contains a label for the sentence's target [91, 117].

- ***Kaggle*[21] *[36, 91].*** This dataset was published by Kaggle for one of Kaggle's competitions. The data used for this competition is social media data and contains binary labels on insult [36, 91].

[15] https://github.com/fhamborg/NewsMTSC
[16] https://github.com/t-davidson/hate-speech-and-offensive-language
[17] http://github.com/zeerakw/hatespeech
[18] http://github.com/zeerakw/hatespeech
[19] https://github.com/Vicomtech/hate-speech-dataset
[20] https://competitions.codalab.org/competitions/19935
[21] https://www.kaggle.com/c/detecting-insults-in-social-commentary

2.4.4 *Summary: Existing Datasets*

In sum, a wide range of datasets for the three tasks exists but with the drawback that annotations are not uniform. The respective labels are often against the background of various media bias types for media bias. For example, the dataset of Baumer, et al. [12] refers to framing bias, whereas the dataset of Chen, et al. [25] aims to detect more detailed types of media bias. Some researchers even completely adjust the dataset labels to match their respective research goals, for example, by labeling sentences as pro-Russian, pro-Western, or neutral [28]. A similar problem occurs for sentiment datasets, as some datasets label the sentiment polarity (cf. [58, 117]), whereas others measure the strength of the polarity (cf. [129]). Unsurprisingly, the same ambiguity occurs for hate speech datasets, as some data are labeled for hate, others for insult or aggression.

Also, the label target is not always comparable, as in some datasets the labels are on statement-level (cf. [33, 58, 108]), for others its on article-level (cf. [89]), and in some cases the annotations are even target-related (cf. [64]).

2.5 FILLING THE GAP

What has been discussed so far is that media bias is a complex phenomenon that can manifest in various ways. In Section 2.1 the most common definitions of media bias are presented, and the differences between individual bias types are explained. It becomes clear that no universal definition of media bias exists, and often, researchers refer to only certain types of media bias. In general, the existing literature provides valuable insights. However, many scholars do not consider the bigger picture when conducting their research. Consequently, it is hard to summarize the current state-of-the-art research.

This lack of a theoretical framework is a significant drawback of media bias research. This thesis aims to fill this gap by defining a first-of-its-kind media bias framework **(C1)**. This framework has already been presented in Section 2.2 and visualized in Figure 2.1. It divides media bias into four categories and assigns different bias types to these categories. In addition, the framework allows better embedding of media bias within the context of other related concepts, which eventually allows a better understanding of the concept of media bias and potentially guides future research more concretely.

Based on the conducted literature review, it has been observed that the majority of media bias research focuses on explaining

- why bias occurs in the news,
- how to detect bias in a statement, an article, or on reporting level, and
- why news consumers perceive news as biased at all.

As already explained in Section 2.2.5, hate speech, and sentiment analysis are two concepts of high informational value for media bias research. The existing literature shows that all three concepts are highly connected. Sentiment Analysis is related to media bias when assuming that a polarized sentence (positive or negative) is considered to be non-neutral and hence biased [72]. Only a few sentiment analysis approaches exist against the background of media bias. Here, the scholars attempt to observe the sentiment of newspaper articles and, by that, obtain insights on the bias level of these articles [45, 67].

In addition, sentiment analysis techniques can be used to detect hateful language [2, 112]. As shown in Section 2.2.5, the majority of research on hate speech focuses on the detection of hateful language on social media platforms. These approaches are usually not related to media bias research. In general, not much work exists that connects hate speech with news articles [143]. In their work, Zannettou, et al. [143] examined this very subject by studying user comments posted on news articles. The authors provide valuable insights on what factors influence the occurrence of hateful comments [143].

In sum, there is work that combines sentiment analysis and research on hate speech detection. Then there is work that combines sentiment analysis and media bias research. Lastly, some work exists that combines hate speech and media bias research. However, to the best of my knowledge, no research exists yet that combines all three concepts. This thesis aims to fill this gap and connect all three approaches.

The base idea of the approach of this thesis is similar to the work of Zannettou, et al. [143] but adapts accordingly to match the research objectives of this project. The overall goal of the analysis is to examine characteristics of user comments in terms of sentiment and hate (henceforth called comment characteristics) and put these comment characteristics in relation to the bias of the respective article. In more detail, this thesis observes whether there are significant differences between the comment characteristics of articles that are more biased compared to those of less biased articles. Hence, it is examined whether the comment section of an article can be an indicator of the article's level of bias.

Lastly, as illustrated in Section 2.4, no dataset suitable for this approach does yet exist. Therefore, the first step is to create a new media bias dataset that comprises hate and sentiment values on statement-level and bias scores on article-level. In conclusion, the three major contributions of this work are:

C1: The development of a universal theoretical framework for media bias detection - the media bias framework (cf. Section 2.2).

C2: The construction of a first-of-its-kind dataset useful for a combined study of media bias, sentiment analysis, and hate speech.

C3: Conducting an analytical study by observing comment characteristics on news articles in order to detect indicators of the article's bias.

2.6 research objectives

Given the media bias framework in Figure 2.1, two implications can be derived: 1) hateful language might be an indicator for bias, and 2) a statement's polarity (positive or negative) might be an indicator for bias. Assuming these implications hold true, then user comments that contain hate or sentiment values can be considered as being biased. From this, the following hypotheses can be derived:

H1: The more hateful the comments on an article, the more biased this article is.

H2: The stronger the comments' polarity on an article, the more biased this article is.

Additionally, research on the perception of media bias shows that individuals that are confronted with biased news tend to adopt similar biased views [65, 83]. Hence, the implication can be drawn that the occurrence of bias in one place might trigger the occurrence of bias in another place. As news outlets exist that clearly represent a political ideology and thus influence news consumers [35], it is justified to additionally state the following hypotheses:

H3: The more biased a news outlet, the more biased are the articles of that news outlet.

This leads to the following two research question:

RQ1: Are user comments on a news article an indicator of the article's bias?

RQ2: Is the new outlets' stance an indicator of the article's bias?

In the following, each step of the following approach and the motivation behind it are outlined. Each part is discussed in more detail in Chapter 3.

The first and most crucial part of this work is to collect suitable data in order to create a high-quality dataset. First, article-related data, i.e., the data that identifies the bias of an article, is collected. This data is accessed via ad fontes media, a corporation that "rate[s] the news for reliability and bias to help people navigate the news landscape" [76]. On ad fontes media's website, a list of articles is provided that have been manually labeled according to their level of bias and reliability. The bias score defines how politically influenced an article is, where the values range from -42 (most extreme left) to +42 (most extreme right). The reliability score, on the other hand, indicates how much truthfulness the article contains. Here, the values range from 0 (least reliable, contains inaccurate/fabricated info) to 64 (most reliable, original fact reporting). The existing literature shows that labels provided by ad fontes media have been used for other media bias-related tasks before and are high-quality [25]. In this thesis, the subject of interest is the article's bias, whereby this does not necessarily have to be the political bias. According to how media bias is defined in Section 2.2, both metrics, bias score and reliability score, are valuable for this research. Additionally, ad fontes media also provides overall bias scores and overall reliability scores for each news outlet. With regard to the second research question stated above, these two metrics are also considered to be of importance.

In the next step, it is required to collect the comments made on the rated articles. Here, several options exist how to collect this user-generated data., for example, directly from the news website or from social media platforms like Twitter or Facebook. Several reasons speak for Twitter as a data source. First, ad fontes media provides article ratings of roughly 300 outlets. One option is to collect the user comments on these 300 news websites directly. However, collecting data from 300 different websites is very time- and resource-intensive. Hence, a more efficient option is to collect user comments from only one website, e.g., a social media website. Second, the nature of social media platforms like Twitter or Facebook allows rapid sharing of personal thoughts, opinions, or information that users are willing to share voluntarily [41]. This user-generated, up-to-date content covers a large variety of topics [94] and is therefore considered a valuable data source for all kinds of text processing tasks, inter alia, sentiment analysis [41, 147] and hate speech detection [119]. Lastly, Twitter is one of the most popular micro-blogging sites [41] with roughly 436 million active users[22]. In a survey of the Reuters Institute, 25% of all Twitter users worldwide stated that they use Twitter to get the latest news [102]. Although 25% might not seem that much, in comparison with other popular Social Media platforms, Twitter is the most popular for news consumption [102]. Following this reasoning, it is concluded that Twitter is a suitable data source for this project. The Twitter API[23] is used to collect the relevant tweets.

Once the data collection process is completed, the article comments then need to be analyzed in order to capture the comment characteristics. The comments are examined with regard to multiple features. First, a transfer learning method is applied in order to detect the sentiment polarity of the tweets. In the next step, this approach is replicated to observe whether the comments contain hateful language. Existing literature shows that a wide range of sentiment analysis and hate speech detection methods exist. As demonstrated in Section 2.3, the use of deep neural network models for text classification has become increasingly popular. One of these methods is XLNet, "a generalized autoregressive pretraining method that [...] enables learning bidirectional contexts" [140, p. 1]. XLNet "overcomes the limitations of BERT thanks to its autoregressive formulation" [140, p. 1] and "integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining" [140, p. 1]. Yang, et al. [140] demonstrated, that

[22] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[23] https://developer.twitter.com/en/products/twitter-api

XLNet "outperforms BERT on 20 tasks, [...] including [...] sentiment analysis" [140, p. 1]. Based on these results, XLNet is considered to be an appropriate method for both of the above-mentioned text classification tasks.

Lastly, the tweets are additionally classified using Google's free API, Perspective API[24], which "uses machine learning models to identify abusive comments" [79]. The Perspective API provides a total of 16 attributes (i.e., emotional concepts) and specifies how likely the respective comment is perceived as said attribute. For example, if a comment has a score of 0.94 for the attribute "toxicity", then the probability that the comment is perceived as toxic is 94%.

After examining the comments with regard to their characteristics, the last step deals with examining the data with regard to the above-stated hypotheses. This eventually allows drawing conclusions that answer the research questions of this paper. To adequately evaluate the data, a multi-level model is applied. Recall that the data collected in the preceding data collection process contains information on article-level as well as on outlet-level. Hence, the data is of hierarchical nature. This means the articles are nested within news outlets. In such cases, multi-level models are recommended, as they "allow [...] to examine the influence of individual (i.e., Level 1) and cluster-level (i.e., Level 2) covariates" [44, p. 121].

[24] https://www.perspectiveapi.com/

3 METHODOLOGY

In the previous Chapter 2, the term media bias has been defined based on a detailed literature review. In addition, a theoretical framework for media bias has been presented, which divides media bias into four categories. Furthermore, state-of-the-art approaches for the detection of media bias, sentiment analysis, and hate speech detection have been presented, as well as a listing of the available datasets. Based on the existing gaps in the media bias research, the research objectives of this thesis have been derived.

This Chapter 3 describes the methodological approach followed throughout this project. Section 3.1 describes the data collection process; Section 3.2 describes how the collected comments are analyzed for their characteristics; Section 3.3 presents the statistical models used to examine the underlying relationships of the collected metrics

3.1 DATA COLLECTION

In this Section 3.1 the data collection process is presented. As described in Section 2.6, data on statement-level (i.e., user-comments on articles) and data on article-level (i.e., the bias of the articles) is required. First, the data on the upper level is collected, which means all articles and their respective bias and reliability scores that have been rated by ad fontes media

The data on statement level is gathered in a second step, based on the list of articles collected in step one. For reasons already stated in Section 2.6, the data is collected from Twitter, using the Twitter API. In the following, both steps of the data collection process are described in more detail.

3.1.1 *Collecting News Articles from ad fontes media*

ad fontes media defines themselves as "a public benefit corporation with a mission to make news consumers smarter and news media better" [76]. For each article ad fontes media has analyzed, they provide a bias score and a reliability score. The bias scores range from -42 (most extreme left) to +42 (most extreme right). The reliability scores range between 0 (least reliable) to 64 (most reliable). In addition, ad fontes media provides overall labels for each outlet considered. These overall metrics consist of the outlet's overall bias score and the outlet's respective bias class, as well as the outlet's overall reliability score and the outlet's respective reliability class. The value ranges for the overall scores are similar to the article-related scores. The bias category consists of seven classes: 1) most extreme left, 2) hyper-partisan left, 3) skews left, 4) middle or balanced bias, 5) skews right, 6) hyper-partisan right, and 7) most extreme right [76]. The reliability category consists of eight classes: 1) original fact reporting, 2) fact reporting, 3) complex analysis or mix of fact reporting and analysis, 4) analysis or high variation in reliability, 5) opinion or high variation in reliability, 6) selective or incomplete story/unfair persuasion/propaganda, 7) contains misleading information, and 8) contains inaccurate/fabricated information [76].

To collect this information, the R package "rvest"[1] is used to scrape the relevant information provided on ad fontes media's website *adfontesmedia.com*[2]. First, a list of all the news sources that have been rated is compiled. From this list, all outlets which received overall ratings (henceforth called relevant outlets), but of which no separate articles have been ranked, are manually excluded. Given this list of relevant outlets, the following article-related metrics are scraped: article headline, article URL, bias score of the article, and reliability score of the article. All article-related metrics are merged into one *.csv* file, together with an outlet identifier (i.e., the URL pointing to the dedicated outlet page on *adfontesmedia.com*). Lastly, some of the article headlines are corrected manually, as the information embedded on *adfontesmedia.com* is not always correct. For example, in some cases, the article headline is only partial, or the article URL is set as a headline.

In addition, for each outlet, the following outlet-related metrics are scraped: overall bias score, overall reliability score, bias class, reliability class, and the proper outlet's name. All outlet-related metrics are merged into a separate *.csv* file.

3.1.2 *Collecting Tweets with the Twitter API*

In the next step, the statement-level data is collected using Twitter as the data source. To access Twitter data, the Twitter API[3] is used. The collection of the tweets happens in several steps, which are gradually explained in the following. As set out in Section 2.6, it is required to find user comments that can uniquely be related to one of the articles rated by ad fontes media. Hence, the first step is to find the respective outlet's tweets that reference one of the rated articles (henceforth called original tweets). Once the original tweets are identified, the comments on these tweets are collected.

Step 1: Manual Preparation

Some manual preparation is required in order to automate the search process. Based on the data collected from ad fontes media, first, the Twitter username, i.e., Twitter handle,

[1] https://cran.r-project.org/web/packages/rvest/rvest.pdf
[2] https://adfontesmedia.com/rankings-by-individual-news-source/; accessed on 2021-10-26
[3] https://developer.twitter.com/en/products/twitter-api

for each news outlet is collected. The list of these Twitter handles is later required as one of the input parameters for the API request.

Second, all publication dates for the articles are manually collected. This data has not been embedded on ad fontes media's website. Hence, it is not possible to automatically gather this information. The articles' publication dates are necessary to define an individual time period for each outlet. These outlet-specific time periods, consisting of a start date and an end date, are later used to specify the API request further. When formulating a request to the Twitter API, several additional query parameters can be added. Two of these additional parameters are `start_time` and `end_time`. Specifying values for these two parameters limit the request to the specific time frame, as only values within this time period are returned by the API.

Hence, in a third step, these time periods are identified, which are necessary for the following reason. This Twitter data collection process is based on finding the original tweets. Based on the article-related data collected from ad fontes media, the only two existing cues that link the original tweets to the articles are the article's headline and the article's URL. However, formulating an API request by setting the query search parameter equal to either one of them would not be very successful. First, because the API request returns only exact string matches, no fuzzy string matching can be defined. Second, the articles' headlines or the articles' URLs are often not contained in the tweet text itself. As a result, it is difficult to specify the API request so that only original tweets are returned, and the chances are high that some original tweets stay uncovered.

For this reason, searching the original tweets happens in two steps. First, all tweets posted by the relevant outlets within a generous time period are collected. Once all potentially relevant tweets are collected, this data is further processed using more sophisticated techniques in the Python environment. However, the time spans of the rated articles are very large with the earliest article date being 2010-05-23, the most recent one being 2021-11-08. Therefore, it would be very inefficient, time- and resource-wise, to scrape all tweets for all outlets within this large time period. Practically, this would mean to scrape all tweets for all outlets within the last ten years. Logically, this is not feasible, which is why an individual time period is determined for each outlet. The procedure is the following: From the list of rated articles, all articles for which no publishing date has been found are removed and articles published before 2019-01-01. These are only a small fraction of articles that are considered as outliers. Next, the publishing dates are grouped by outlets, and the minimum date and the maximum date are specified. The start date is then defined as three days prior to the minimum date (e.g., if the minimum date for the outlet is 2021-05-04, the start date is set to 2021-05-01). The end date is by setting the maximum date seven days forward (e.g., if the maximum date for the outlet is 2021-10-01, the end date is set to 2021-10-08). These time shifts are made to adjust for potential deviations from the article's publishing date (e.g., if the article was published on the outlet's website on 2021-07-02, but the tweet is posted only two days later). Additionally, a cut-off date is defined in order to prevent future dates, as the API request cannot process these. This is necessary in cases where the publishing date is less than seven days prior to the execution day.

Step 2: Searching the Original Tweets

Once the manual preprocessing is completed, all API request parameters are prepared adequately. Hence, the next step addresses the issue of finding the original tweets.

First, all tweets of all relevant outlets within the previously defined time periods are collected. This is done within a Python environment by implementing a for-loop, iterating over the list of Twitter handles, start dates, and end dates. The parameters of the API request are defined as the following:

- `'query': f'from:{handle}'`
- `'start_time': datetime.datetime.strptime(start,'%Y-%m-%dT%H:%M:%S%z'). isoformat()`
- `'end_time': datetime.datetime.strptime(end,'%Y-%m-%dT%H:%M:%S%z'). isoformat()`
- `'tweet.fields':'author_id, entities, attachments, conversation_id, created_at, referenced_tweets'`
- `'max_results': 500`

The first parameter, `query`, is the only required parameter. By specifying the value for `query` defines what the request returns. In this case, the `query` is defined as `from:handle` which returns all tweets from the desired Twitter user (i.e., the outlet). Here, `handle` functions as a place holder while iterating over the list of all outlets' Twitter handles. For example, to collect all tweets posted by the news outlet 19th News, the value for `query` must be specified as `from:19thnews`, as 19thnews is the respective Twitter handles of that outlet. The two following parameters `start_time` and `end_time` limit the request on a temporal level. If a start date and an end date are specified, only tweets that have been posted within this time period are returned. The necessity for start and end dates has already been stated above. The next request parameter, `tweet.fields`, specifies which Tweet fields are returned by the API request. These are defined as `author_id, entities, attachments, conversation_id, created_at, referenced_tweets`, of which especially `author_id` and `entities` are relevant for the next steps. Lastly, `max_results` defines the maximum number of tweets returned with one API request. This number is set to `500` as it is the maximum possible number. In addition to the request parameters, the endpoint URL needs to be defined. In order to search tweets, the following URL must be provided `https://api.twitter.com/2/tweets/search/all`.

Once all potentially relevant tweets have been collected, the data is further processed by loading it into a Python environment. In order to find the relevant tweets, a sequence of regex-based matching is applied. Here, the focus is on finding tweets based on the article's URL, which is the only unique identifier. Among other things, the API request returns the fields `entities`, which for each tweet contains several meta-information. This nested meta-information includes two entries called `expanded_url` and `unwound_url`, which both safe the destination URL in case the tweet itself included a link to an external page. It is not entirely clear when the URL is anchored as `expanded_url`, and when as `unwound_url`. Thus, in order to find as many original tweets as possible, both parameters are checked.

In addition, not all collected tweets contain information about `expanded_url` and `unwound_url`. Unfortunately, this means that URL matching is not always successful, and the risk is high that many original tweets stay uncovered. However, observing the collected Twitter data reveals that some outlets partially or fully replicate the article headline within the tweet. Hence, a third option is to find original tweets by scanning the tweet text for the article headline.

Hence, the three matching steps are:

1. Search if the article URL matches `expanded_url` embedded in the meta-information. Return all tweets for which a match is found.

2. For all tweets for which no match has been found in the first step, repeat the process by checking if the article URL matches the `unwound_url`. Again, return all tweets for which a match is found.

3. Lastly, in case both preceding matching steps have not been successful, try to find original tweets by scanning the tweet texts for the articles' headlines.

Once all collected tweets have been scanned, a list of all original tweets is compiled, i.e., all tweets that reference one of the articles rated by ad fontes media.

Step 3: Collecting Comments on Original Tweets

After working out the list of original tweets in step two, the comments on the articles are collected in the next step. First, all original tweets that have no comments are filtered out from the list. This is done for efficiency reasons. The Twitter API has the limitation that only 300 requests can be made within a 15 minutes time frame. In order to not exceed the time limit, a time out of four seconds is defined, which means that one request is made every 4th second. For example, if the list of original tweets contains 7000 articles, the process of scraping all comments of these 7000 articles takes at least roughly 8 hours. However, if 3000 of these articles have not been commented on, they can be filtered out in the first place. Hence, the run time improves to last only 4.5 hours. By using the endpoint URL `http://api.twitter.com/2/tweets/counts/all`, an API request can be specified that returns a count of all comments made on a tweet. The parameters for the API request are defined as the following:

- `'query': f'conversation_id:{tweet_id}'`
- `'start_time': datetime.datetime.strptime(start,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'end_time': datetime.datetime.strptime(end,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'granularity': 'day`

To understand this API request, one must know Twitter's specification behind the parameters `id` and `conversation_id`. Each tweet has a unique id, by which tweets are distinguished. In addition, all tweets within the same thread have an identical conversation id that, in turn, is equal to the original tweet's id. Consequently, one can find all comments to a tweet by matching the parameters `id` and `conversation_id`. Hence, this time the `query` is specified to return all tweets where the conversation id is equal to the original tweet's id. Additionally, a value for the parameters `start_date` and `end_date` is provided which are different to the time periods defined in Step two. Here, the `start_date` is equal to the date of the original tweet. The `end_date` is created by setting the tweet date 60 days forward. The duration of 60 days is considered as an adequate time frame, based on experience from previous work [20, 121]. Most of the times where researchers collected Twitter data, a time period of 2 months has been considered [20, 121]. Lastly, `granularityj` defines how the counts are grouped, which can be per minute, per hour, or per day.

Once the API request has returned the counts for all original tweets, all twets with zero comments are filtered out. The API request to collect all comments is then specified, using the following parameters.

- `'query': f'conversation_id:{tweet_id}'`
- `'start_time': datetime.datetime.strptime(start,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'end_time': datetime.datetime.strptime(end,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'tweet.fields':'in_reply_to_user_id, author_id, created_at, conversation_id'`
- `'expansions': 'referenced_tweets.id, in_reply_to_user_id'`
- `'max_results': 500`

The first three parameters, `query`, `start_date`, and `end_date`, are specified similar as in the preceding step. These are the most important parameters here. Additioanlly, the parameters `tweet.fields` and `expansions`, are specified. Again, `max_results` is set to the maximum possible number, which is 500. Similar as to the tweet search conducted in step 2, the endpoint URL is again set to `https://api.twitter.com/2/tweets/search/all`.

Lastly, the same procedure is conducted for retweets. More specifically, Twitter distinguishes between retweets and quoted retweets. Retweets are shared posts where one user retweets another user's post without adding text themselves. Quoted retweets, on the other hand, allow the user to add their own texts. For this project, only quoted retweets are considered. The API request is similar to the above, with only one minor difference in the parameters:

- `'query': f'{tweet_id}'`
- `'start_time': datetime.datetime.strptime(start,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'end_time': datetime.datetime.strptime(end,'%Y-%m-%dT%H:%M:%S%z').isoformat()`
- `'tweet.fields':'in_reply_to_user_id, author_id, created_at, conversation_id'`
- `'expansions': 'referenced_tweets.id, in_reply_to_user_id'`
- `'max_results': 500`

The only difference in the API request is how the parameter `query` is defined. By specifying it as shown above, the API request returns all retweets, including simple retweets and quoted retweets. However, the returned data contains an attribute called `type` that states whether the retweet is quoted or not. Only retweets where `type` equals `quoted` are kept.

3.1.3 *Compiling the Dataset*

After collecting the relevant data, all information is compiled into one dataset. In short, the dataset consists of two parts: 1) article-related data, and 2) outlet-related data. The article-related data have been scraped from ad fontes media. It contain information about the bias and the reliability of particular articles. Additional article-related data have been collected from Twitter, using the Twitter API. This data contains comments on tweets, which in turn reference the articles rated by ad fontes media. The outlet-related data have also been scraped from ad fontes media and contain ratings about the overall bias and overall reliability of particular news outlet. The dataset is described in more detail in Section 4.1.

3.2 EXAMINING COMMENT CHARACTERISTICS

As already described in Section 2.6, after the data collection process is completed, the collected comments are examined concerning their characteristics. This process is separated into two parts: 1) the transfer learning method in order to examine the comments' sentiment polarity and hatefulness, and 2) the classification approach using Google's Perspective API[4].

For reasons already stated in Section 2.6, the pretrained model of the XLNet method[5] [140] is adopted, which is fine-tuned once for sentiment analysis and once for hate speech detection respectively. Both learning procedures rely on a similar model set up, however, the subtle differences are explained in detail in the following Section 3.2.1. The entire fine-tuning procedure is implemented[6] on Kaggle[7], using Kaggle's free access to GPU.

In Section 3.2.2, it is explained how the tweets are classified using Google's Perspective API.

3.2.1 *Transfer Learning: Fine-Tuning XLNet for Text Classification*

The XLNet is "a generalized autoregressive pretraining method" [140, p. 1], that improves current state-of-the-art techniques on text classification by combining the best of two of the most successful pretraining methods, autoregressive and autoencoding language modeling [140]. The general idea behind XLNet is, to predict a word's probability by capturing bidirectional contexts. This means, in contrast to commons language models, the XLNet is not "using a fixed forward or backward factorization order" [140, p. 2] but "maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order" [140, p. 2]. In other words, the XLNet determines "the context for each position [consisting] of tokens from both left and right" [140, p. 2]. Hence, XLNet considers interdependencies between words, which is contrasting BERT's independent assumption, stating that "tokens are [predicted] independent of each other" [140, p. 2].

Comparing XLNet to other existing language models shows, that XLNET outperforms even established language models like BERT and RoBERTa for a range of NLP-related tasks, including sentiment analysis and text classification tasks [140]. Given this well-founded neural architecture of XLNet and its state-of-the-art results, XLNet is

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

[5] https://huggingface.co/xlnet-base-cased

[6] The implementation is based on this article and code example:
https://medium.com/swlh/using-xlnet-for-sentiment-classification-cfa948e65e85; accessed on 2022-01-05
https://github.com/shanayghag/Sentiment-classification-using-XLNet; accessed on 2022-01-05

[7] https://www.kaggle.com/

considered a suitable method for further examining the comment characteristics with regard to sentiment polarity and hatefulness.

The pretrained model is accessed using the Hugging Face's library. There, two pretrained XLNet models are available, the `xlnet-base-cased`, and the `xlnet-large-cased`. Both are applicable for English language, however, the base model is a 12-layer, 768-hidden, 12-heads, 110M parameters architecture, whereas the large model is a 24-layer, 1024-hidden, 16-heads, 340M parameters[8]. Because of limited computational resources, the smaller `xlnet-base-cased` is used.

XLNet for Sentiment Classification

The fine-tuning procedure is split into two main tasks: 1) choosing appropriate training data and 2) defining suitable model parameters.

In order to fine-tune the `xlnet-base-cased`, labeled training data is required. As shown in Section 2.4.2, several publicly available sentiment datasets exist. Hence, for time and efficiency reasons and because sufficient amounts of labeled data for sentiment analysis are available, it is refrained from manually labeling new data. Choosing a good training dataset is highly important, as the classification results can only be as good as the training data. Hence, if the training data is labeled poorly, the classifier will most likely not provide good classification results for new data. For this task, the Stanford Twitter Sentiment dataset [58], Sentiment140[9], is used. The reasons for that are the following. On the one hand, Sentiment140 also contains Twitter data which means that the data structure is similar. On the other hand, the dataset contains roughly 1.6 million tweets labeled for sentiment polarity (positive, neutral, and negative), which means that the amount of available data is certainly large enough.

The dataset is available via Hugging Faces library[10] and is split into two parts: a training set and a test set with the test set being manually labeled [58]. Manually annotated data usually indicates high-quality labels. However, the test set contains only 498 tweets. Unfortunately, this is not sufficient to fine-tune the sentiment analysis model. Therefore, the training set is used. Additionally, the number of researchers who used Sentiment140 for sentiment analysis speak for using this dataset [10, 58, 147].

The training set of Sentiment140 accessed via Hugging Face's library contains labels for positive sentiment (4) and negative sentiment (0). In order to keep labels uniform, the labels are adjusted so that negative sentiment is denoted by 0 and positive sentiment is denoted by 1. For fine-tuning the XLNet, a smaller subset is created by random sampling 24,000 of the 1.6 million tweets. The class distribution is roughly equal, with 12,017 tweets being labeled as negative, and 11,983 as positive (cf. Figure 3.1).

Before the data is further used, the tweet texts are preprocessed. This includes removing text passages that contain no cues for the sentiment of a statement, for

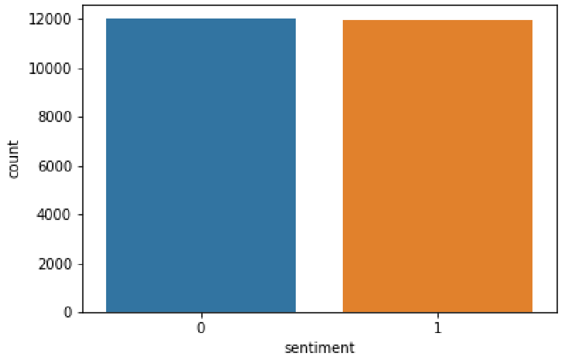
[8] https://huggingface.co/transformers/v2.0.0/pretrained_models.html

[9] http://help.sentiment140.com/for-students

[10] https://huggingface.co/datasets/sentiment140

example, a Twitter user's handle (i.e., "@user"), URLs, or other unnecessary characters like hash signs, multiple whitespaces, or tabs. Additionally, smileys are removed, as they cannot be adequately processed.

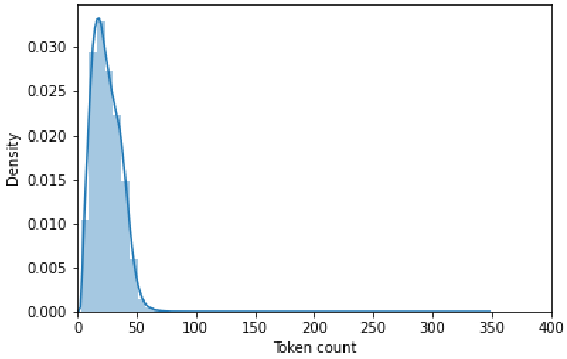
Figure 3.1: Distribution of Sentiment Class Labels in the Training Data



Note: This Figure shows the distribution of class labels for the dataset used for fine-tuning XLNet for sentiment analysis. The total dataset size is 24,000 tweets, of which 12,017 tweets are labeled as negative (0), 11,983 tweets are labeled as positive (1).

The second part of the fine-tuning process addresses choosing the correct hyperparameters for the model set up. As the specification of the model parameters strongly influences the resulting model, this step is crucial in order to obtain the desired classification model, providing acceptable results.

Figure 3.2: Distribution of Token Length in Sentiment Dataset



Note: This Figure shows the distribution of the input data length, i.e., the number of tokens per tweet, for the sentiment dataset used for fine-tuning XLNet.

First, the two hyperparameters MAX_LEN and BATCH_SIZE are defined. MAX_LEN defines the maximum length of the input tokens. In this case, the length of an input token is the number of tokens per tweet. Due to the architecture of XLNet, the maximum possible value is 512. Hence, inputs that are longer than 512 tokens are truncated. In Figure 3.2 the distribution of the tweet lengths over the entire dataset are visualized. The Figure shows that although most tweets have a token length between 0 and 100, some samples of the input data are longer. In order to avoid truncation, MAX_LEN is set to 400.

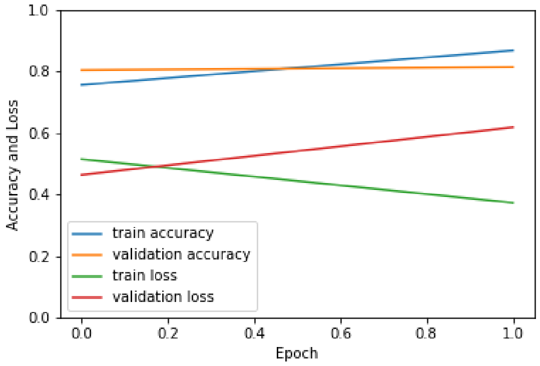
BATCH_SIZE, on the other hand, defines how many samples are trained in one iteration. The higher this value, the more data samples are processed in one iteration, thus, the faster the model executes training. However, a higher BATCH_SIZE also requires higher computational power. Due to limited computational resources, the value for BATCH_SIZE is set to 8.

In the next step, the input data is split into three parts: a training set (50%), a test set (25%), and a validation test (25%). The training set is used to initially train the model, which is then validated with the validation set in order to get the model's performance after each training epoch. The final model is then tested using the test set to obtain unbiased estimates for the model's fit.

Additional hyperparameters that need to be defined are the optimization function and the dropout rate. The optimization function is defined as AdamW [82] with a learning rate of $3e^{-5}$. The dropout rate is specified within the pretrained model as 10%. In general, the "idea of dropout is to randomly drop units and relevant connections from neural networks during training" [141, p. 4]. This has the purpose to "prevent units from co-adapting too much" [141, p. 4] and hence is an "effective technique against overfitting" [141, p. 4]. In this case, the dropout rate is specified within the pretrained model as 10%.

Lastly, the number of training epochs is defined. One epoch is completed when all samples of the training and validation set have been run through the model once. If the number of epochs is not chosen carefully, the trained model might suffer from under- or overfitting. Underfitting is the case when the model fails to remember the trained data structures. In contrast, overfitting occurs when the model remembers the data structure too well and thus, is not generalizable enough [141]. Both under- and overfitting lead to the model performing poorly and thus should be avoided. To determine whether the model suffers from over- or underfitting, one should keep an eye on the learning curves during the training process. Training loss and validation loss are cues for the model's quality. If the training loss is larger than the validation loss, then the model suffers from underfitting, whereas the opposite is true if the training loss is smaller than the validation loss. Ideally, the model should be trained until a point where the validation loss exceeds the training loss in order to avoid underfitting. At the same time, the model should not be trained too long, as the more the validation loss exceeds the training loss, the more the model suffers from overfitting.

Figure 3.3: Learning Curves for Fine-Tuning XLNet for Sentiment Analysis



Note: This Figure shows the loss and accuracy for the training and validation process over time. The number of epochs is set to 2, where x=0.0 reflects the values after the first epoch, x=1.0 the values after the second epoch respectively.

Having a look at the learning curves displayed in Figure 3.3 shows that after one training epoch (x=0.0), the model is not sufficiently trained yet. Consequently, one epoch is not enough. However, after two epochs (x=1.0), the validation loss exceeds the training loss, and training should be stopped. For this reason, the value for `EPOCHS` is set to 2.

Table 3.1: Classification Report: XLNet for Sentiment Analysis

	Precision	Recall	F1-Score	Support
Negative	0.820	0.815	0.818	3023
Positive	0.814	0.818	0.816	2977
Accuracy			0.818	6000
Macro Average	0.817	0.817	0.817	6000
Weighted Average	0.817	0.817	0.817	6000

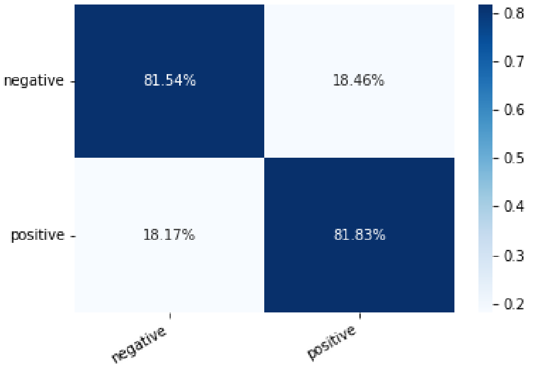
Note: This Table presents the results from the classification report obtained by fine-tuning XLNet for sentiment analysis.

After fine-tuning the model, the final model is eventually evaluated using new data (i.e., test data). The classification report in Table 3.1 provides information on how well the final model performs on new data. The metrics of the classification report are explained in the following:

- Precision: "[T]he precision for a class is the number of instances correctly labeled as belonging to that class with respect to the total number of elements labeled as belonging to that class [14, p. 107]. In other words, precision is "a measure of exactness" [14, p. 107].
- Recall: The "[r]ecall is the number of correctly classified instances with respect to the total number of objects belonging to that class [14, p. 107]. In other words, recall is "a measure of completeness" [14, p. 107].
- F1-Score: The F1-score is " a weighted average of precision and recall, where the F-measure reaches its best value at 1 and its worst score at 0" [14, p. 108]
- Support: The absolute number of instances present in the dataset, belonging to the respective class [14].

With regard to the classification results shown in Table 3.1, it can be concluded that the fine-tuned `xlnet-base-cased` provides a decent performance. The overall F1-score of 81.8% is good, although there is still room for improvement. The classification report shows that the model performs slightly better at detecting negative sentiment than positive sentiment. The precision for negative sentiment states that 82% of instances that have been classified as negative are correctly classified. For positive sentiment, this value is 81.4%.

Figure 3.4: Confusion Matrix: XLNet for Sentiment Analysis



Note: This Figure shows the confusion matrix for the evaluated model. The classification of negative sentiment and positive sentiment is almost identical. However, for both classes, roughly 20% are wrongly classified.

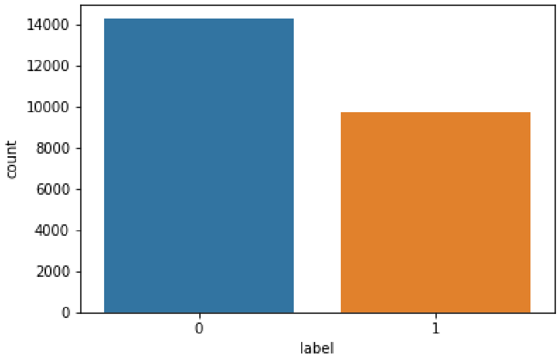
Lastly, the confusion matrix in Figure 3.4 shows that for both classes, the proportion of true negative and true positive (dark blue cells) is similar and is approximately 80%. Reversely, this means that for both classes, almost 20% of the predictions are wrong.

In conclusion, the fine-tuned model for sentiment analysis provides acceptable results, however, they could be better. The learning curves in Figure 3.3 already anticipated that the model suffers from overfitting. Several methods exist how to avoid overfitting, for example, by using different datasets or by further adjusting the hyperparameters. Due to time constraints and the fact that the model still works adequately, this was not further pursued. However, this issue is discussed in more detail in Section 5.1.

XLNet for Hate Speech Detection

The implementation of the fine-tuning of `xlnet-base-cased` for hate speech detection follows a similar procedure than explained for sentiment analysis.

Figure 3.5: Distribution of Hate Class Labels in the Training Data



Note: This Figure shows the distribution of class labels for the dataset used for fine-tuning XLNet for hate speech detection. The total dataset size is 24,000 tweets, of which 14,279 tweets are labeled as non-hate (0), 9,721 tweets are labeled as hate (1).

Similar to sentiment analysis, labeled hate speech data is required. In order to fine-tune `xlnet-base-cased` for hate speech detection, the HatebaseTwitter dataset [33] is used. This dataset is a collection of tweets that have been labeled as hate speech, offensive language, or neither. For the purposes of this project, it is not distinguished between hateful language and offensive language. According to how hate speech is defined in Section 2.2, the existence of both hateful language and offensive language is considered to be biased language. However, the HatebaseTwitter dataset [33] available via Huggin Face's library[11], is highly imbalanced, with 19,190 tweets labeled as offensive, 1,430 tweets labeled as hateful, and only 4,163 tweets labeled as non-hate. For this reason, another dataset has been added, also available via Hugging Face's library[12],

[11] https://huggingface.co/datasets/hate_speech_offensive

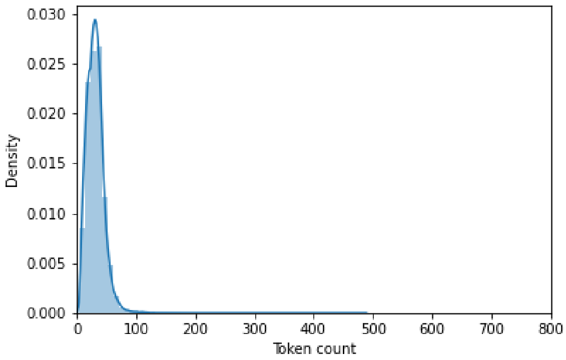
[12] https://huggingface.co/datasets/tweets_hate_speech_detection

which contains additional 31,962 tweets, of which 29,720 tweets are labeled as neutral, and 2,242 tweets are labeled as hateful.

To create the dataset used for fine-tuning, both hate datasets are combined, and then 24,000 tweets are randomly sampled, of which 9,721 are labeled as hate, and 14,279 labeled as non-hate (Figure 3.5). The tweets are again preprocessed for further usage. The preprocessing is the same as already described for the sentiment dataset above.

Most of the hyperparameters are defined the same as already for the fine-tuning of the XLNet for Sentiment Analysis. These are the `BATCH_SIZE`, which is again set to 8 due to computational limitations, the dropout rate of the pretrained model is still specified as 10%, and again AdamW is used as an optimization function with a learning rate of $3e^{-5}$. In addition, the input data is partitioned in the same ratio, i.e., the training set is 50%, and both validation set and test set is 25% each.

Figure 3.6: Distribution of Token Length in Hate Dataset



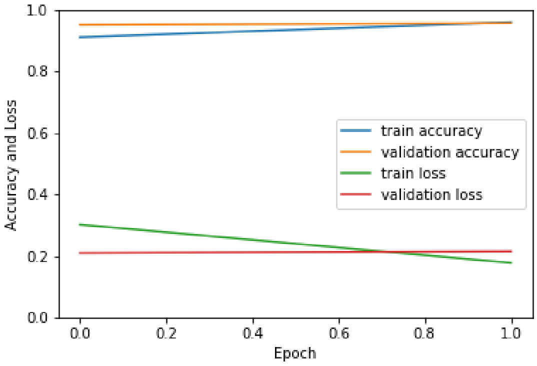
Note: This Figure shows the distribution of the input data length, i.e., the number of tokens per tweet, for the hate speech dataset used for fine-tuning XLNet.

The only difference is the parameter `MAX_LEN`, which is this time set to 512. Figure 3.6 shows the distribution of the tweet lengths for the hate dataset, which clearly shows that some instances of the dataset have a length of almost 500, which justifies setting `MAX_LEN` to 512.

The number of epochs is set to 2. Having a look at the learning curves in Figure 3.7 shows the loss and accuracy after the first epoch (x=0.0) and after the second epoch (x=1.0). It can be concluded that one epoch is not sufficient for training as the training loss is still larger than the validation loss. However, after the second epoch, the validation loss exceeds the training loss. Hence, setting `EPOCH` to 2 is considered to be appropriate.

The classification report in Table 3.2 shows, that the fine-tuned `xlnet-base-cased` for hate speech detection states good results. The overall F1-score of the model is 95.5%, which is significantly higher compared to the sentiment analysis model. The

Figure 3.7: Learning Curves for Fine-Tuning XLNet for Hate Speech Detection



Note: This Figure shows the loss and accuracy for the training and validation process over time. The number of epochs is set to 2, where x=0.0 reflects the values after the first epoch, x=1.0 the values after the second epoch respectively.

classification report also shows that the model's predictions are slightly better for tweets labeled as no-hate than for tweets labeled as hate.

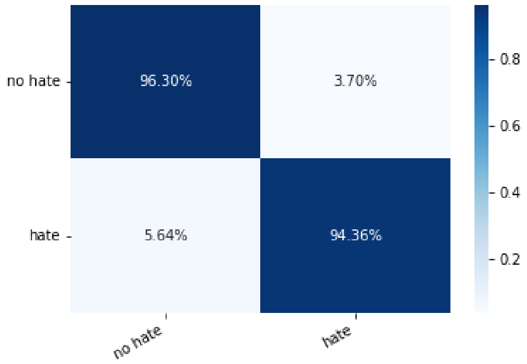
Table 3.2: Classification Report: XLnet for Hate Speech Detection

	Precision	Recall	F1-Score	Support
No Hate	0.962	0.963	0.962	3569
Hate	0.946	0.944	0.945	2431
Accuracy			0.955	6000
Macro Average	0.954	0.953	0.953	6000
Weighted Average	0.955	0.955	0.955	6000

Note: This Table presents the results from the classification report obtained by fine-tuning XLNet for hate speech detection.

Lastly, the confusion matrix in Figure 3.8 shows that for both classes, the proportion of false predictions is small. Roughly 96.30% of no-hate instances have been correctly classified, 94.36% of hate instances respectively.

Figure 3.8: Confusion Matrix: XLNet for Hate Speech Detection



Note: This Figure shows the confusion matrix for the evaluated model. The classification of hate and no-hate is almost identical The proportions of wrongly predicted instances is only small.

In conclusion, the fine-tuned model for hate speech detection provides good results. Compared to the fine-tuned model for sentiment analysis, this model for hate speech detection provides better classification results. The learning curves in Figure 3.7 show that the model hardly suffers from overfitting. Hence, it is no surprise that the model fit is better for this classification task.

3.2.2 *Perspective API*

The Perspective API is a free API "hosted on Google Cloud Platform" [79]. It uses machine learning techniques in order to classify text into multiple attributes, i.e., emotional concepts. By formulating the API request respectively, it returns the requested attributes' scores for a particular text. The scores range between 0 and 1 and indicate the likelihood that a reader perceives the text as containing said attribute. For example, if the attribute score for "toxicity" is 0.75, the probability that the text is perceived as toxic is 75%. It is possible to choose from a number of attributes and individually request only scores for chosen attributes. There are 16 attributes in total, which are explained in more detail in Section A.1.1. The scores for all 16 attributes have been requested. Unlike hateful language and sentiment polarity, not all of the 16 attributes are forms of bias. However, they still represent comment characteristics and thus might be indicators for bias.

Before the attribute scores are requested, the tweets are again preprocessed in the same way as described above for the two XLNet-based fine-tuning tasks.

The API request contains only a few parameters, which are specified as the following:

- `'text': tweet`
- `'requestedAttributes': 'TOXICITY', 'SEVERE_TOXICITY', 'THREAT', 'IDENTITY_ATTACK', 'INSULT', 'PROFANITY', 'SEXUALLY_EXPLICIT', 'FLIRTATION', 'ATTACK_ON_AUTHOR', 'ATTACK_ON_COMMENTER', 'INCOHERENT', 'INFLAMMATORY', 'LIKELY_TO_REJECT', 'OBSCENE', 'SPAM', 'UNSUBSTANTIAL'`
- `'languages': 'en'`

The first parameter, `text`, defines the input for which the attributes are requested. With `requestedAttributes`, one can specify which attributes are returned. In this case, it includes all 16 attributes. Lastly, `languages` defines the language of the input text. In this case, this is only English. However, some of the attributes also work for other languages like German, Spanish, or French. The endpoint URL is set to `https://commentanalyzer.googleapis.com/$discovery/rest?version=v1alpha1`.

3.2.3 *Summary: Examining Comment Characteristics*

The key points that have been discussed in this Section 3.2 are how the collected tweets are examined with respect to their characteristics. Two methods have been presented: 1) text classification using a pretrained language model, which has been fine-tuned once for the subtask sentiment analysis, once for hate speech detection, respectively, and 2) text classification using Google's Perspective API. Once this step of examining the comment characteristics is completed, the final dataset contains all information required to observe the hypotheses stated in Section 2.6. The method for this last step is described in the following Section 3.3.

3.3 MULTI-LEVEL MODELING

Multi-level models are statistical models that enable the examination of hierarchical data structures. This means data that is distributed on two or more hierarchical levels. A typical example for such nested data is observations of students nested in different schools [44]. However, this is only a simple example, and multi-level models can deal with even more complex data, for example, residents belonging to a city, which belongs to a county, which in turn belongs to a country. According to Hox, et al. [71], "[a] multi-level problem is a problem that concerns the relationships between variables that are measured at a number of different hierarchical levels" [71, p. 4]. When dealing with hierarchical data but using, for example, a standard multiple regression analysis, i.e., "analyz[ing] all available data at one single level" [71, p. 4], leads to "conceptional and statistical problems" [71, p. 4]. Multi-level models are useful for examining, for example, "how a number of individual and group variables influence one single individual outcome variable" [71, p. 4]. In other words, with multi-level models, one can "examine the influence of individual (i.e., Level 1) and cluster-level (i.e., Level 2) covariates" [44, p. 121].

3.3.1 *The 2-Level Regression Model*

In the following, the basic model set up for a 2-level regression model is explained, including two level-1 predictors (X_1 and X_2) and two level-2 predictors (Z_1 and Z_2). The peculiarity of such a hierarchical data structure is that the level-1 predictors usually have variance on both levels, which can be of different impact [44, 71]. The variance on level 1 is referred to as within-group variance, the variance on level 2 as between-group variance respectively [44]. In contrast, level-2 predictors typically have only variance on the upper-level [44, 71].

The level-1 regression Equation with two level-1 variables (X_1 and X_2) looks as follows [44, 71]:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + e_{ij} \tag{3.1}$$

In Equation 3.1, β_{0j} is the intercept for group j, β_{1j} the regression coefficient of X_{1ij} for group j, β_{2j} the regression coefficient of X_{2ij} for group j, and e_{ij} the level-1 residual error term [44, 71]. The subscript j determines the group, the subscript i the group member respectively.

The variation of the intercept β_{0j} and the regression coefficients β_{1j} and β_{2j} from Equation 3.1 are dependent from the two level-2 explanatory variables [71]:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} \tag{3.2}$$

Equation 3.2 predicts the average value for Y in group j, depending on the two level-2 explanatory variables Z_1 and Z_2 [71]. Hence, if γ_{01} is positive, the average value of Y in group j is higher for a higher value of Z_{1j} [71]. If γ_{02} is positive, the same applies for Z_{2j} respectively.

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + u_{1j} \tag{3.3}$$

Equation 3.3 determines, that the relationship between the outcome Y and the level-1 explanatory variable X_1 depends on the two level-2 explanatory variables Z_1 and Z_2 [71]. Hence, if γ_{11} and γ_{12} are positive, then the effect of X_1 on the outcome Y is stronger with higher values for Z_1 and Z_2 respectively [71].

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Z_{1j} + \gamma_{22}Z_{2j} + u_{2j} \tag{3.4}$$

For Equation 3.4 the same is true with regard to X_2 respectively. Note that the terms $\gamma_{11}Z_{1j}X_{1ij}$, $\gamma_{12}Z_{2j}X_{1ij}$, $\gamma_{21}Z_{1j}X_{2ij}$, and $\gamma_{22}Z_{2j}X_{2ij}$ in Equation 3.3 and Equation 3.4 indicate the interaction effects. For all three Equations, the u indicates the residual error at class level [71].

The combined model regression Equation is now obtained by substituting Equation 3.2, Equation 3.3, and Equation 3.4 into Equation 3.1 [44]:

$$\begin{aligned} Y_{ij} = \gamma_{00} &+ \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} \\ &+ \gamma_{10}X_{1ij} + \gamma_{11}Z_{1j}X_{1ij} + \gamma_{12}Z_{2j}X_{1ij} \\ &+ \gamma_{20}X_{2ij} + \gamma_{21}Z_{1j}X_{2ij} + \gamma_{22}Z_{2j}X_{2ij} \\ &+ u_{0j} + u_{1j}X_{1ij} + u_{2j}X_{2ij} + e_{ij} \end{aligned} \tag{3.5}$$

The two terms $\gamma_{01}Z_{1j}$ and $\gamma_{02}Z_{2j}$ in the first line describe the effects of the level-2 predictors Z_{1j} and Z_{2j} respectively on the outcome Y. The second and third lines of the Equation represent the effects of the level-1 predictors X_{1ij} and X_{2ij} on the outcome Y. Note, that these effects are potentially moderated by the level-2 predictors. Lastly, the fourth line of the Equation combines all residual errors at class level.

3.3.2 *The Pure Effects of Level-1 and Level-2 Predictors*

However, in order to obtain interpretable parameter estimates, especially against the background of the research questions, it is necessary to center the parameters. Although centering is widely used in ordinary least squares regression, in the context of multilevel models, it is more complex [44]. In the following, grand mean centering and group mean centering are introduced.

When centering around the grand mean, then for each value of the variable, the overall mean is subtracted, which in turn produces a new value, henceforth denoted by the subscript CGM [44]. Hence, the formula to calculate the grand mean centered variable is

$$X_{ij_{CGM}} = X_{ij} - \overline{X}, \tag{3.6}$$

where $\overline{X}$ is the mean over all observations.

In contrast, when centering around the group mean, then for each value of the variable, its respective group mean is subtracted, which in turn produces a new variable, henceforth denoted by the subscript CWC [44]. The group mean is calculated as the mean over all values belonging to group j. In this case, the formula to calculate the group mean centered variable is

$$X_{ij_{CWC}} = X_{ij} - \overline{X}_j, \tag{3.7}$$

where $\overline{X}_j$ is the mean over all observations belonging to group j.

In addition, for variables that have variance on both hierarchical levels, the obtained group means can additionally be grand mean-centered. By that, an additional level-2 variable is created that models the variance of level-1 variables on the upper level. The formula for that is

$$X_{j_{CGM}} = \overline{X}_j - \overline{X}, \tag{3.8}$$

where $\overline{X}_j$ is the group mean of group j and $\overline{X}$ the grand mean.

After that being said, whether a variable should be group mean centered or grand mean centered highly depends on the question of interest. In general, level-1 variables can be centered either around the group mean or around the grand mean [44]. Level-2 variables, on the other hand, can only be centered around the group mean as they do not have variance on the lower level [44].

Keeping in mind the two research questions stated in Section 2.6, the following objectives are of interest: 1) the effect of comment characteristics on the article's bias, and 2) the influence of the outlet's characteristics on the article's bias. This means the effects of level-1 variables (i.e., hatefulness and sentiment polarity of the comments) on Y as well as the effects of level-2 variables (i.e., the outlet's overall bias and overall reliability) on Y are observed. Ideally, the "pure" effects of the level-1 and level-2 predictors are desired in order to provide meaningful interpretations. Hence, the following centering is applied:

- Group mean centering of the two level-1 predictors hate score and sentiment polarity.
- Grand mean centering of the two level-2 predictors overall bias and overall reliability.
- Grand mean centering of the two level-1 group means for hate score and sentiment polarity.

Hence, the final multi-level regression model will look as follows. For simplicity reasons, interaction effects are not included for now, as these are only interesting, in case a significant relationship between level-1 predictors and Y is observed [44].

$$\begin{aligned} Y_{ij} = \gamma_{00} &+ \gamma_{01} Z_{1j_{CGM}} + \gamma_{02} Z_{2j_{CGM}} + \gamma_{03} Z_{3j_{CGM}} + \gamma_{04} Z_{4j_{CGM}} \\ &+ \gamma_{10} X_{1ij_{CWC}} + \gamma_{20} X_{2ij_{CWC}} \\ &+ u_{0j} + u_{1j} X_{1ij} + u_{2j} X_{2ij} + e_{ij} \end{aligned} \tag{3.9}$$

The regression coefficients of Equation 3.9 can be interpreted as follows:

- γ_{00}: The mean intercept.
- γ_{01} & γ_{02}: Predict the "pure" effects of the level-2 predictors Z_1 (γ_{01}) and Z_2 (γ_{02}) on the outcome variable. For example, if γ_{01} is positive, the outcome variable is higher when the value for Z_1 is higher.

- γ_{10} & γ_{20}: Predict the "pure" effects of the level-1 predictors X_1 (γ_{10}) and X_2 (γ_{20}) on the outcome variable. For example, if γ_{10} is positive, the outcome variable is higher when the value for X_1 is larger. Hence, these two regression coefficients indicate the within-group variance.
- γ_{03} & γ_{04}: Predict the variance of the two level-1 predictors hate score and sentiment polarity on level-2. Hence, these two regression coefficients indicate the between-group variance.

3.3.3 *Interaction Effects of Level-1 and Level-2 Predictors*

If the results of the multi-level regression model from Equation 3.9 provide significant results for level-1 relationships, the model can be adjusted accordingly. In case both level-1 relationships are significant, i.e., the relationship between hate score and Y, and the relationship between sentiment polarity and Y, then the model will look as shown in the following:

$$
\begin{aligned}
Y_{ij} = \gamma_{00} &+ \gamma_{01} Z_{1j_{CGM}} + \gamma_{02} Z_{2j_{CGM}} \\
&+ \gamma_{10} X_{1ij_{CWC}} + \gamma_{11} Z_{1j_{CGM}} X_{1ij_{CWC}} + \gamma_{12} Z_{2j_{CGM}} X_{1ij_{CWC}} \\
&+ \gamma_{20} X_{2ij_{CWC}} + \gamma_{21} Z_{1j_{CGM}} X_{2ij_{CWC}} + \gamma_{22} Z_{2j_{CGM}} X_{2ij_{CWC}} \\
&+ u_{0j} + u_{1j} X_{1ij} + u_{2j} X_{2ij} + e_{ij}
\end{aligned}
\tag{3.10}
$$

In this case, the focus lies on the four regression coefficients γ_{11}, γ_{12}, γ_{21}, and γ_{22}, which predict the respective interaction effects.

4 RESULTS

The preceding Chapter 3 describes the underlying methodology of this project, which can be divided into three major parts: 1) the data collection process, 2) the examination of the comment characteristics, and 3) the multi-level regression model in order to investigate the two the main question of this thesis. First, the results for the web scraping procedure and the Twitter data collection step are provided in Section 4.1. In Section 4.2 the classification results are summarized, using the three classification approaches described in Section 3.2. Lastly, in Section 4.4, the results obtained from the multi-level regression analysis are presented. The three proposed hypotheses are tested, which eventually provide answers to the research question of this thesis.

4.1 DATA COLLECTION

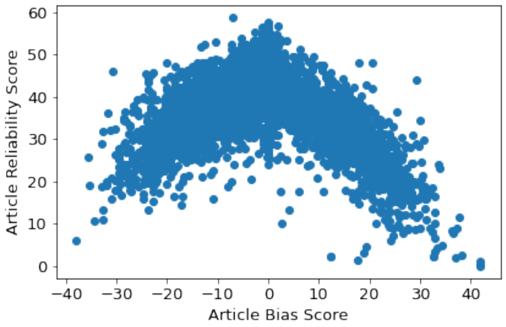
4.1.1 *Data Collected from ad fontes media*

In total, ad fontes media provides outlet-related information about 321 news outlets. As the focus lies on the article-related metrics in the first place, only those outlets have been considered where at least one article has been rated. Hence, the data collection process as described in Section 3.1.1 results in a dataset containing a total of 6,345 rated news articles of 283 different news outlets.

The majority of articles have rather low bias scores, with most articles being centered around a bias score of 0 (i.e., not politically biased). However, there are slightly more left-skewed articles than right-skewed articles. Overall, the articles are predominantly rated as reliable, with only a few articles having low reliability scores. A similar pattern is observed for the bias and reliability scores of the outlets. The majority of rated outlets are left-biased with fewer right-skewed outlets. Overall, the outlets are mostly considered to be reliable. In Section A.2, Figure A.1, Figure A.2, Figure A.3, and Figure A.4 show the plots for the above-described distribution of bias and reliability scores.

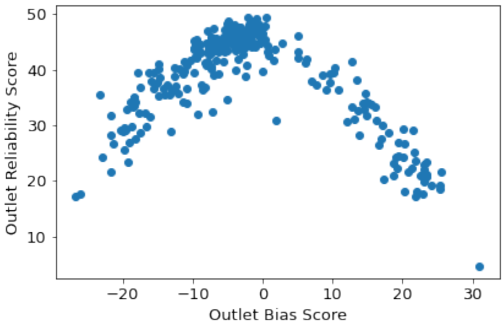
Interestingly, the following Figure 4.1 indicates a relationship between the level of bias of an article and its level of reliability. From the plot, the conclusion is derived that more biased articles are also less reliable. Figure 4.2 shows the same relationship on outlet-level. Both plots show a similar pattern, which in turn strengthens the hypothesis stated in Section 2.6, that the outlet's level of bias influences its articles' bias (H3).

Figure 4.1: Bias Score vs. Reliability Score for Articles



Note: This Figure shows the pattern between the political bias of articles and their reliability. It is observable, that more biased articles tend to have lower reliability scores.

Figure 4.2: Bias Score vs. Reliability Score for Outlets



Note: This Figure shows the pattern between the political bias of outlets and their reliability. It is observable, that more biased outlets tend to have lower reliability scores.

4.1.2 *Data Collected from Twitter*

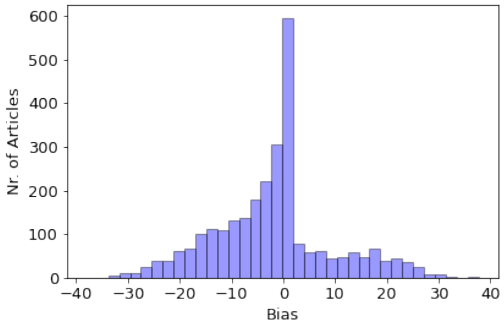
Based on the articles collected from ad fontes media, the next step is to find the original tweets, i.e., the tweets referencing one of the rated articles. Keeping in mind the procedure described in Section 3.1.2, first, the Twitter handles for all outlets have been collected, and an individual time period for each outlet has been defined.

After scraping all tweets of the 283 targeted outlets and after the sequence of regex-based matching has been applied, a total number of 7,059 original tweets have been found. However, some of these tweets link to the same article. This is the case, for example, when the outlet posts a tweet referencing a news story multiple times. In total, from the 6,345 articles collected from ad fontes media, only 3,473 articles of 268 outlets remain. Consequently, for 15 outlets, no original tweets have been found. The reasons for that are either that no tweets have been posted within the dedicated time period, or no tweets referencing the respective articles have been posted, or the matching process was unsuccessful. A detailed overview of included and excluded outlets is listed in Table A.2.

In the next step, all comments on these 7,059 original tweets are collected. As already explained in Section 3.1.2, comments that are directly posted below the original tweet, as well as the quoted retweets, are considered. This last step of the Twitter data collection process results in a dataset containing a total number of 175,807 comments and quoted retweets. Henceforth, all direct comments and quoted retweets are referred to as comments, as the distinction between direct comments and quoted retweets will no longer be necessary. These 175,807 collected comments refer to 2,800 articles of 255 news outlets. Not all of the original tweets have been commented on, hence the decrease in the number of articles. Additionally, 13 outlets have been entirely excluded as no comments on the original tweets have been collected.

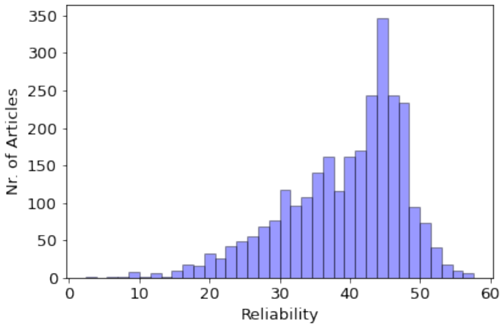
During the Twitter data collection process, roughly two-third of the data accessible at ad fontes media's website have been excluded. The reasons for that are that either no original tweets have been found or the original tweet has not been commented on. The plots below show the distribution of bias and reliability scores for all articles and outlets included in the final dataset. Overall, the amount of left- and right-biased articles, shown in Figure 4.3, is roughly similar as in the initial data. The same applies to the reliability levels of the articles, shown in Figure 4.4, although some of the articles having medium to high reliability have been removed. In contrast, the distribution plot for the bias scores among all outlets, displayed in Figure 4.5, as well as the plot for the distribution of reliability scores, shown in Figure 4.6, have not changed significantly. Hence, removing the respective articles and outlets during the Twitter data collection process did not change the underlying structure of the data.

Figure 4.3: Distribution of Bias Scores among all Articles



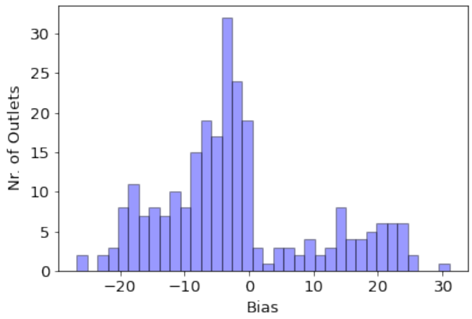
Note: This Figure shows the distribution of the articles' bias scores over the dataset, where the bias ranges from -42 (hyperpartisan left) to +42 (hyperpartisan right).

Figure 4.4: Distribution of Reliability Scores among all Articles



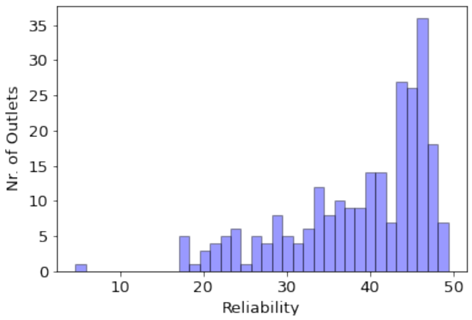
Note: This Figure shows the distribution of the articles' reliability scores over the dataset, where the score ranges from 0 (most unreliable) to +64 (most reliable).

Figure 4.5: Distribution of Bias Scores among all Outlets



Note: This Figure shows the distribution of the outlets' bias scores over the dataset, where the bias ranges from -42 (hyperpartisan left) to +42 (hyperpartisan right).

Figure 4.6: Distribution of Reliability Scores among all Outlets



Note: This Figure shows the distribution of the outlets' reliability scores over the dataset, where the score ranges from 0 (most unreliable) to +64 (most reliable).

4.2 EXAMINING COMMENT CHARACTERISTICS: CLASSIFICATION RESULTS

4.2.1 *XLNet for Sentiment Analysis*

Using the fine-tuned `xlnet-base-cased` for sentiment analysis obtained in Section 3.2.1, the tweets are classified according to their sentiment polarity. The classifier returns two scores, one for the positive sentiment and one for the negative sentiment. Both scores range between 0 and 1, where the positive sentiment score indicates the likelihood of the tweet being positive, and the negative sentiment score indicates the likelihood of the tweet being negative. In sum, both scores add up to 1, which means the scores are reciprocal. Given both scores, the tweets are then eventually labeled as either "positive" or "negative", depending on which score is larger. An example is provided in the following Table 4.1.

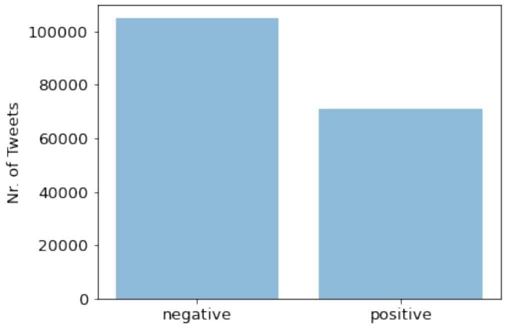
Of all 175,807 tweets in the dataset, 59.53% have been classified as negative, 40.47% of the tweets as positive. The plot is shown in Figure 4.7. In addition, Figure 4.8 shows the distribution of the sentiment score, where a value close to 0 refers to negative polarity, a value close to 1 to positive polarity. The Figure shows that most results are quite explicit, being either close to 0 or close to 1.

Table 4.1: Classification Results using XLNet for Sentiment Analysis

Tweet Text	Positive Score	Negative Score	Label
Cool Never going back to work maskedforever	0.163	0.837	negative
Excellent choice as is I feel so lucky as a Californian to have such amazing representation	0.994	0.006	positive

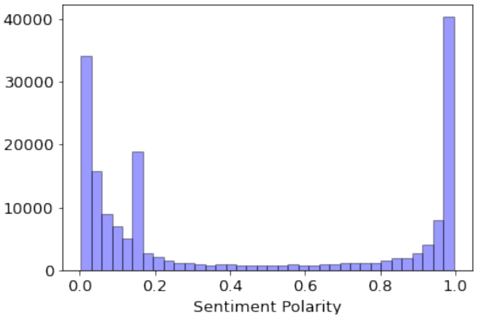
Note: This Table provides examples, how the fine-tuned `xlnet-base-cased` obtained in Section 3.2.1 classifies text into positive or negative sentiment. As described in that Section, the tweet text has been cleaned for a better text understanding.

Figure 4.7: Distribution of Class Labels for Sentiment Analysis



Note: This Figure shows the distribution of the sentiment labels determined using the XLNet-based classifier. 59.53% of the tweets are labeled as negative, 40.47% as positive.

Figure 4.8: Distribution of Sentiment Scores



Note: This Figure shows the distribution of the sentiment polarity scores. Values close to 0 refer to negative polarity, values close to 1 to positive polarity.

Given the sentiment scores, a third sentiment-related attribute is added to the dataset, which captures the strength of the polarity independent of its direction. One of the hypotheses stated in Section 2.6 says that the stronger a comment's sentiment polarity, the more biased the article (H3). Since the sentiment scores lie between 0 and 1, with both ends of the range indicating negative and positive sentiment, respectively, the actual polarity strength is not given by that. For example, considering a negative sentiment score of 0.95 simultaneously means that the positive sentiment score is 0.05. Hence, the statement is classified as negative because the negative sentiment score exceeds the positive score. However, that score does not indicate how strong the sentiment is, i.e., how much it deviates from a neutral value. The value for neutral sentiment is 0.5, as it is the exact middle value between the two extremes. Therefore, in order to obtain the polarity strength, the absolute differences for each sentiment score to 0.5 are considered. Because of the scores' mutuality, it does not make a difference which score to take. The absolute distances to 0.5 are the same for both positive and negative scores. Lastly, to keep this attribute on a comparable scale with the other comment characteristics, the polarity strengths are transformed onto a scale ranging from 0 to 1, using min-max normalization.

4.2.2 *XLNet for Hate Speech Detection*

As described in Section 3.2.1, a hate speech classifier has been implemented with the goal to detect hateful language in tweets. For this purpose, tweets are classified as either hate or non-hate. Similar to the sentiment classifier, this classifier also returns two scores, one indicating the likelihood that the text is hateful and the other indicating the appositive. Again, both scores range between 0 (non-hate) and 1 (hate) and add up to 1. Examples for the hate speech classification results are provided in Table 4.2.

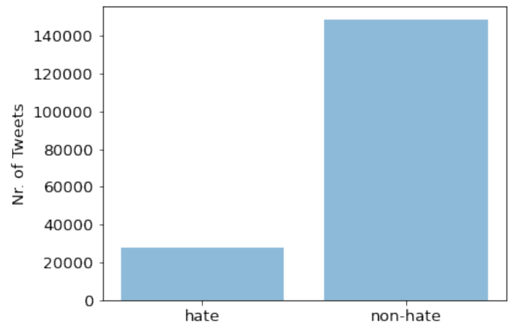
Table 4.2: Classification Results using XLNet for Hate Speech Detection

Tweet Text	Positive Score	Negative Score	Label
Because she is a better person than I am	0.016	0.984	non-hate
let s see if he listens or does the same pigheaded shit he did with USDA	0.985	0.015	hate

Note: This Table provides examples, how the fine-tuned `xlnet-base-cased` obtained in Section 3.2.1 classifies text into hate and non-hate. As described in that Section, the tweet text has been cleaned for a better text understanding.

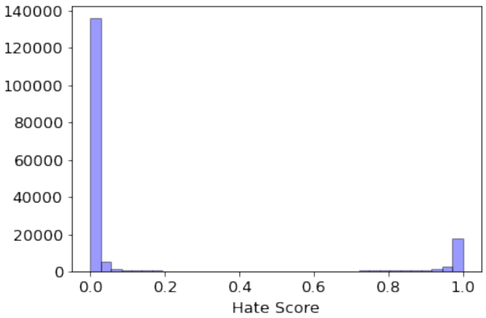
Of all 175,807 tweets in the dataset, only 15.7% have been classified as hate, whereas the vast majority of 84.3% tweets have been classified as non-hate. The plot is shown in Figure 4.9. Interestingly, the classifier provides more stable results than the one for sentiment analysis. As shown in Figure 4.10, the hate scores are even more explicit with values being either close to 0 (non-hate) or close to 1 (hate).

Figure 4.9: Distribution of Class Labels for Hate Speech Detection



Note: This Figure shows the distribution of the hate labels determined using the XLNet-based classifier. 15.7% of the tweets are labeled as hate, 84.3% as non-hate.

Figure 4.10: Distribution of Hate Scores



Note: This Figure shows the distribution of the hate values. Values close to 0 refer to non-hate, values close to 1 to hate.

4.2.3 *Perspective API for Text Classification*

The last step described in Section 3.2.2 refers to text classification using Google's Perspective API. The ratings for all existing 16 attributes have been requested. The API then returns a score between 0 and 1, indicating the likelihood that the respective tweet is perceived as said attribute. Table 4.3 provides an example for three selected attributes: "toxicity", "profanity", and "insult". The attribute toxicity describes "[a] rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" [79]. Profanity means "[s]wear words, curse words, or other obscene or profane language" [79]. A statement having a high score for insult means, it is an [i]nsulting, inflammatory, or negative comments towards a person or a group of people" [79]. A detailed description of all attributes is provided in Table A.1.

Table 4.3: Classification Results using Perspective API

Tweet Text	Toxicity	Profanity	Insult
Trump is a cheap cheating lying bastard	0.981	0.956	0.985
I don t know I like Ellen Always have	0.129	0.088	0.057

Note: This Table provides examples of how Google's Perspective API classifies text into the three attributes "toxicity", "profanity", and "insult". As described in that Section, the tweet text has been cleaned for a better text understanding.

Most of the scores follow a similar distribution throughout the dataset, where scores are either skewed towards 0 or towards 1. Only the two attributes, "incoherent" and "inflammatory", have more balanced scores throughout the dataset. In Section A.2, Figure A.5 shows all distribution plots for the 16 Perspective API attributes.

4.3 THE FINAL DATASET

Before continuing with the results of the multi-level regression model, this Section shortly summarizes the data that has been obtained.

The final dataset consists of a total number of 175,807 tweets on 2,800 articles. Overall, 255 news outlets have been included in the dataset. In total, the dataset can be divided into three types of attributes: 1) comment characteristics, 2) article-related metrics, and 3) outlet-related metrics. The comment characteristics refer to all the attributes that determine the tweets' characteristics: the positive and negative hate scores, the hate value, the positive and negative sentiment scores, the sentiment value, the polarity strength, and the 16 Perspective API attributes. Second, article-related metrics refer to all information on article-level that have been collected from ad fontes media. Those attributes are the bias score and the reliability score. Lastly, outlet-related metrics have also been collected from ad fontes media. They are the outlet's overall bias score and respective bias class, and the outlet's overall reliability score and its reliability class. A detailed overview of all attributes of the dataset is found in Table A.3.

4.4 MULTI-LEVEL REGRESSION ANALYSIS

Before the data is further analyzed, it has first been observed by plotting a correlation matrix. However, the results are not meaningful because they show only weak relations between the attributes in question (cf. Figure A.6, Figure A.6). This most likely stems from the reason that the underlying structure of the dataset corresponds to hierarchical data. More precisely, the created dataset contains information on tweet-level, article-level, and outlet-level. The tweets are nested within articles, and the articles are in turn nested within outlets. However, the focus lies on the bias of an article, which is reflected by the two attributes bias score and reliability score. As explanatory factors, all attributes describing comment characteristics as well as the outlet scores are considered. Because the dependent variables are on article-level, it is assumed to be the lowest level (i.e., level 1) and outlets the upper level (i.e., level 2). Consequently, all tweet-related comment characteristics are grouped by articles. By that, the comment characteristics are transformed to level-1 variables. This makes the regression analysis a multi-level problem on two hierarchical levels.

The following multi-level regression is conducted only with a subset of the available attributes, which are considered to be most meaningful. The data included consists of two level-1 predictors, two level-2 predictors, and two outcome variables. The predictors on level 1 are hate score and polarity strength. The positive hate score is included for the hate score, which is interpreted as the higher the score, the more hateful the comment. On the other hand, the polarity strength indicates the higher the score, the stronger the sentences' polarity. However, if the polarity is positive or negative is not included in this score. The predictors on level 2 are the outlet's overall political bias and overall reliability. The two outcome variables are also on level 1 and refer to the article's bias score and the article's reliability score. As both outcome variables are of interest, the regression analysis is conducted once for bias score as the dependent variable and once for reliability score as the dependent variable.

Before the multi-level regression model can be performed, the two bias scores need to be prepared in order to obtain meaningful results. With the biases ranging from -42 to +42, the interpretation of regression results will be difficult, as the exact direction cannot be uniquely identified. Therefore, all bias scores are transformed onto the positive scale, which leads to the two attributes ranging from 0 to 42, where 0 indicates no bias and 42 indicates most extreme biased. By that, the direction of the effects can be unambiguously interpreted.

4.4.1 *The "Pure" Effects of Level-1 and Level-2 Predictors*

Recall the multi-level regression Equation stated in Equation 3.9, which is used to obtain the "pure" effects of the predictors. The model consists of two level-1 predictors and two level-2 predictors. In order to obtain meaningful results with interpretable regression coefficients, the predictors are centered accordingly. Table 4.4 provides an overview of the model parameters, the regression coefficients, the corresponding predictors of the model, and which centering method has been applied.

As stated above, the multi-level regression is conducted once for the articles' bias scores as the outcome variable and once for the article's reliability score, respectively. Table 4.5 shows the parameter estimates for two models, where (1) refers to the model set up with the bias score as the dependent variable, and (2) refers to the model set up with the reliability score as the dependent variable.

Table 4.4: The Model Parameters used to Estimate the "Pure" Effects

Parameter	Regression Coefficient	Variable	Centering Method
$X_{1_{CWC}}$	γ_{10}	hate_cwc	CWC*
$X_{2_{CWC}}$	γ_{20}	polarity_cwc	CWC*
$Z_{1_{CGM}}$	γ_{01}	overall_bias_cgm	CGM**
$Z_{2_{CGM}}$	γ_{02}	overall_reliability_cgm	CGM**
$Z_{3_{CGM}}$	γ_{03}	gmean_hate_cgm	CGM**
$Z_{4_{CGM}}$	γ_{04}	gmean_polarity_cgm	CGM**

Note: This Table provides a description for the parameters of Equation 3.9.
* Group mean centered
** Grand mean centered

Table 4.5: Results for Level-1 and Level-2 Effects

	Dependent variable:	
	bias_score_abs	reliability_score
	(1)	(2)
hate_cwc	4.065^{***} (0.792)	-2.099^{***} (0.702)
polarity_cwc	0.756 (1.009)	-1.528 (1.036)
overall_bias_cgm	0.793^{***} (0.046)	-0.099^{**} (0.050)
overall_reliability_cgm	-0.057 (0.039)	0.638^{***} (0.043)
gmean_hate_cgm	0.977 (2.182)	-2.953 (2.400)
gmean_polarity_cgm	-0.711 (2.694)	2.087 (3.036)
Constant	8.469^{***} (0.136)	39.252^{***} (0.144)
Observations	2,800	2,800
Log Likelihood	-8,602.005	-9,014.642
Akaike Inf. Crit.	17,232.010	18,047.280
Bayesian Inf. Crit.	17,315.130	18,100.720

Note: *p<0.1; **p<0.05; ***p<0.01

The Effects on the Article Bias

The parameter estimates of the first model provide valuable insights. Starting with the level-1 predictors, the only significant relationship that has been observed is the one between the level of hate in the comments and the article bias. However, this effect refers to the within-group variance. Hence, it can be interpreted as the "pure" effect of the comments' hatefulness on the article's bias. The relationship is positive, and the regression coefficient is 4.065, which means that if the comments' hatefulness increases by 1 score point, the bias of the article increases by 4.065 score points. The "pure" effect of the polarity strength on the article's bias is 0.756. However, the estimate is not significant. In contrast, the between-group variance predicts how a groups' average value affects the average outcome value. These effects are estimated by including the two grand mean centered group means of hate score and polarity strength. The regression coefficients estimating the between-group variance are γ_{03} and γ_{04}. These estimates are not significant.

For the level-2 predictors, the "pure" effect of the outlet's overall bias on the outcome is positive and significant, meaning that the bias score of articles is higher when the outlet is generally considered to be more biased. The effect of the outlet's overall reliability on the outcome is not significant.

In addition, it has been assumed that the slopes for the two level-1 predictors vary across outlets. However, only the variance of the hate score is significant. The variance of the polarity strength is not. This means, for polarity strength, the hypothesis that the slope is varying across outlets can be rejected [71]. Hence, for further research, one can assume that polarity strength is not varying across outlets.

The Effects on the Article Reliability:

The parameter estimates obtained from the second multi-level regression (2) provide similar results as the first regression. As for the level-1 predictors, again, the only significant relationship is observed between the level of hate in the comments and the article's reliability. The value for the regression coefficient is -2.099 and significant at $p < 0.01$, predicting the "pure" effect of hate score on the article's reliability on level 1. This is interpreted as the increase of the comments' hate score by 1 score point, decreases the article's reliability by 2.099 score points. Similar to the results described above, no significant direct effect of polarity strength on the article's reliability has been observed, as well as no significant between-group variances.

For the level-2 predictors, however, both have significant effects on the outcome. For the overall bias, the regression coefficient is -0.099 and significant at $p < 0.05$. This result indicates that the article's reliability decreases with an increase in the outlet's overall bias score. For the overall reliability, the regression coefficient is 0.638 and is significant at $p < 0.01$. This is interpreted as an article's reliability score increasing if the outlet's overall reliability score increases.

In addition, for this model, the variances of the two level-1 predictors are both not significant, allowing the conclusion that the slopes for hate score and polarity strength do not vary across outlets [71]. For further research, this means that both level-1 predictors can be assumed to not vary across outlets [71]

In sum, the two conducted analyses provide evidence that the hate score of the comments has a positive effect on the article's bias. This is true for both outcomes, article bias, and article reliability. In contrast, no significant effects have been observed for the polarity strength. Lastly, the results show significant effects of the outlet's overall bias and the outlet's overall reliability on the article's bias. For the regression on the article's bias score, evidence exists that the article's bias score is higher for outlets rated as more biased. For the regression on the article's reliability score, the parameter estimates indicate that the article's reliability is less for more biased outlets but higher for more reliable outlets.

4.4.2 *Interaction Effects of Level-2 Predictors*

As stated in Section 3.3.3, in the case of significant relationships between level-1 predictors and the outcome variable, the data is further observed with regard to interaction effects. Recalling Equation 3.10, which is used to obtain interaction effects. The model for the interaction effects is specified similarly to the above model, with the only difference that the grand mean centered group means are not required here. The reason for that is, that in order to get estimates for the interaction effects, only the four regression coefficients γ_{11}, γ_{12}, γ_{21}, and γ_{22} are required.

For both above-stated regressions, significant relationships have only been observed between the hate score and the article's bias and the hate score and the article's reliability. Hence Equation 3.10 reduces as shown in the following. The parameters have the same meaning as stated in Table 4.4.

$$\begin{aligned} Y_{ij} = \gamma_{00} &+ \gamma_{01} Z_{1j_{CGM}} + \gamma_{02} Z_{2j_{CGM}} \\ &+ \gamma_{10} X_{1ij_{CWC}} + \gamma_{11} Z_{1j_{CGM}} X_{1ij_{CWC}} + \gamma_{12} Z_{2j_{CGM}} X_{1ij_{CWC}} \\ &+ u_{0j} + u_{1j} X_{1ij} + u_{2j} X_{2ij} + e_{ij} \end{aligned} \tag{4.1}$$

The following Table 4.6 presents the parameter estimates for the interaction effects between hate score and overall bias as well as hate score and overall reliability. The analysis has again been conducted twice: once for bias score as the dependent variable (3) and once for reliability score as the dependent variable (4).

Table 4.6: Results for Interaction Effects

	Dependent variable:	
	bias_score_abs	reliability_score
	(3)	(4)
hate_cwc	3.786*** (0.762)	−2.022*** (0.709)
polarity_cwc	0.684 (1.008)	−1.498 (1.037)
overall_bias_cgm	0.775*** (0.046)	−0.112** (0.049)
overall_reliability_cgm	−0.070* (0.040)	0.634*** (0.042)
hate_cwc:overall_bias_cgm	0.703*** (0.252)	−0.296 (0.235)
hate_cwc:overall_reliability_cgm	0.435** (0.217)	−0.226 (0.202)
Constant	8.461*** (0.136)	39.257*** (0.145)
Observations	2,800	2,800
Log Likelihood	-8,597.700	-9,014.854
Akaike Inf. Crit.	17,223.400	18,047.710
Bayesian Inf. Crit.	17,306.520	18,101.140
Note:		*p<0.1; **p<0.05; ***p<0.01

The Moderating Effect on Hate-Bias Relationship

The parameter estimates for the first regression (3) prove that both interaction effects are significant. The first regression coefficient for the interaction effect between overall bias and hate score is 0.703 and significant at $p < 0.01$. This result indicates that the effect of hateful comments on the article's bias is more prominent for more biased outlets.

The regression coefficient for the interaction effect between overall reliability and hate score is 0.435 and is significant at $p < 0.05$. This is interpreted as the effect of hateful comments on the article's bias is larger for outlets that are considered to be more reliable.

The Moderating Effect on Hate-Reliability Relationship

The parameter estimates for the second regression (4) provide no evidence for the existence of interaction effects. The two regression coefficients for the interaction effect between overall bias and hate score and overall reliability and hate score are not significant.

4.4.3 *Implications from the Regression Results*

Recalling the three hypotheses that have been stated in Section 2.6:

H1: The more hateful the comments on an article, the more biased this article is.

H2: The stronger the comments' polarity on an article, the more biased this article is.

H3: The more biased a news outlet, the more biased are the articles of that news outlet.

In conclusion, the two conducted regression models (1) and (2) support hypothesis 1. Both models provide parameter estimates that show a significant relationship between the hatefulness of comments and the article's bias. In contrast, no evidence has been found that confirms hypothesis 2, indicating that the polarity strength does not affect the article's bias. Lastly, both regression models (1) and (2) show a positive relationship between the outlet's overall bias and the article's bias. In addition, model (2) additionally indicates that the higher the outlet's reliability, the higher the article's reliability. These results confirm hypothesis 3.

In addition, model (3) provides evidence for the existence of interaction effects, indicating that the effect of hateful comments on the article's bias is even worse when the outlet is more biased. These findings underpin the above-described results of the direct effects. Hence, H1 and H3 are supported even stronger.

With regard to the two research questions formulated in Section 2.6, the following conclusion is derived: The above-conducted regression confirms that comment characteristics can indeed be an indicator for the article's bias. However, this has only been observed for the hate score of the comments. The polarity strength seems to have no effect on the article's bias. In addition, evidence has been provided that the outlet's bias also influences how biased the articles are. Hence, the outlets' stance is an additional important factor, next to the comment characteristics.

5 DISCUSSION

This thesis's main goal is to better understand the structure of media bias. As discussed in Chapter 2, the concept of media bias is multi-layered and complex, and current research does not provide a universal definition. By conducting an extensive literature review, some gaps in the current media bias research have been identified, and attempts to fill these gaps have been made. First, one downside of the current state of the art stems from the vast amounts of existing media bias related literature and the fact that no structured theoretical framework of media bias yet exists. Often, researchers provide valuable insights, however, without considering the bigger picture. In order to fill this gap, a theoretical framework of media bias has been proposed in Section 2.2. The framework divides the concept into four subcategories to which different bias types are assigned.

In addition, the existing literature shows that media bias is very closely linked to other concepts, two of which are hate speech and sentiment analysis. Since there is no research yet that explicitly studies these two concepts in the context of media bias, an approach has been proposed to observe whether comments on an article are indicators of the article's bias. For this purpose, first, comments on articles have been collected. Then, these comments have been examined for their characteristics, focusing on hate speech and sentiment analysis. Each comment has been assigned a hate score and a sentiment score, which respectively indicates how much the respective comment contains hate, or in which direction and how strong the sentiment polarity of the comment is. Once these comment characteristics have been determined, it is examined how these characteristics relate to the article's bias they have been posted on. For this last step, multi-level regression models have been applied. Chapter 3 provides an in-depth explanation of the methodological procedure.

Chapter 4 presents the step-wise results of the approach described in Chapter 3. This includes the description of the collected data and presents the results of the comment analysis and the multi-level regression. The results provide evidence that the characteristics of the comments on an article indeed allow making inferences about the article's bias. The regression estimates obtained from several multi-level regressions show that the more hateful the comments, the higher the article's political bias as well as its reliability. However, for the polarity strength, this effect has not been confirmed. In addition, this hate-bias relationship is reinforced by the outlet's overall level of bias and reliability. This indicates that the more the outlet is generally biased and the more the outlet is considered to be reliable, the stronger the effect of hateful comments on the article's bias. These results allow the conclusion that both the comments and the outlet's stance influence the article's bias.

In the following Section 5.1, the limitations of the approach proposed in Chapter 3 are discussed. In addition, Section 5.2 summarizes possibilities for future research.

5.1 LIMITATIONS

One of the limitations of this approach has already been pointed out in Section 3.2.1. As shown in Figure 3.3 the learning curves of the fine-tuning of the XLNet for sentiment analysis indicate that the model suffers from overfitting. Several solutions exists how overfitting can be prevented. The most straightforward solution is to find the appropriate number of training epochs. If the model is trained too long, it remembers the structure of the training data too well and hence results in overfitting [141]. However, here the model suffers from overfitting already after the second epoch. Therefore, a more suitable solution is to increase the quality and the size of the training data, as the model's "performance can be significantly affected by the quantity and quality of training dataset" [141, p. 3]. Ideally, the model is fine-tuned on large amounts of high-quality training data, potentially even applying regularization techniques like experimenting with the dropout rate [141]. However, this approach has not been pursued further due to limited computational resources, time constraints, and the limited number of publicly available sentiment datasets.

In general, the classification of Twitter data with regard to hate speech and sentiment polarity is difficult. This stems from the nature of Twitter as a microblogging service. In general, tweets are short messages with a maximum of only 280 characters. Therefore, users often tend to use abbreviations, smileys, and other Twitter-specific languages to express their opinions [147]. In addition, language models generally have difficulties understanding subtle nuances in human language like negations, sarcasm, or slang.

What is more, that often, the labels of the dataset are not gold-standard. One problem with manually annotated data is that it is prone to contain racial bias or other kinds of biases introduced by the annotator [115]. When training a classifier with biased data, the algorithm will adopt this bias, and hence, the classifier tends to return biased classification results [97].

Therefore, in order to obtain a high-quality, fine-tuned language model for the respective task, a collection of the most common language models should have been considered. By doing so, a baseline performance is established against which the performance of other models can be evaluated, for example, following the approach of Spinde, et al. [125]. In more detail, this means, in addition to the XLNet, other established language models should have been applied. The current state-of-the-art models for text classification include BERT, RoBERTa, DistilBERT, XLM, or T5 [17]. In addition, different approaches can be tested against each other, for example, pursuing the approach presented by Rodrìguez, et al. [112] who applied sentiment and emotion analysis to detect hateful language. It is arguable whether hate speech detection and sentiment analysis are two-class classification problems. Approaches exist where the classification task is considered a multi-class problem. For hate speech detection, some datasets contain three or four labels, specifying the tweets, for example, into hateful language, offensive language, or neither [33, 135, 136]. The same applies for sentiment datasets, where the vast majority of datasets contain at least the three labels positive, negative, and neutral [58, 78].

While collecting the Twitter data, a total number of 28 outlets has been excluded. For 11 of them, the reason is that no original tweets have been found. This observation allows the conclusion that the sequential regex-based matching is not optimized yet. In this work, the approach only checks if the article URL is embedded within the meta-data of the scraped Twitter data. If this is not the case, it is tested whether the tweet text contains the article's headline or parts of it. Obviously, this approach is error-prone. Either wrong matches are returned, especially when the article's headline is rather short and more general (e.g., "temperatures break another heat record"). Or matches are not returned at all, as most tweets do not contain the article's headline. One way to potentially optimize the matching process is to find the articles via the tweet text directly. In most cases, the URLs in the tweet text are short links that do not provide any reference to the original article URL. Therefore, a script is required that automatically clicks each link present in the tweet text and then saves the outputted URL. By that, a collection of all URLs that have been posted is created. This list can then be easily compared to the list of article URLs collected from ad fontes media. This approach is more sophisticated and most probably increases the number of correct matches. However, due to limited time and because the amount of data collected is already adequately high, it has been renounced to consider this approach.

In general, more Twitter data can be collected by recursively collecting comments and quoted retweets. In this thesis, only the comments and quoted retweets on level 1 have been collected. This means only the comments that have been posted directly on the original tweet as well as the quoted retweets of that original tweet. In the following, one could recursively collect all comments to the quoted retweets as well quoted retweets of those retweets.

Lastly, the theoretical framework defined for media bias distinguishes different bias types as adequately as possible, making it easier to understand how individual bias types differ from one another, where similarities exist, and what characteristics can be used to identify the respective bias type. However, due to the complexity of the concept and the complexity of human language, there are always cases where a clear distinction between individual bias types is not possible. For example,

5.2 FUTURE WORK

The discussion of limitations of the approaches proposed in this thesis, some implications for future research have already been made. These are mainly proposals to ensure a better quality of the techniques applied in this thesis. However, this Section 5.2 discusses additional opportunities for future work that go beyond the scope of this thesis.

First, the next logical step is to further observe indicators for an article's bias. Given the hierarchical data structure of this work, additional hierarchies can be added. For example, by conducting topic modeling, a third level can be introduced to the multi-level regression model. This results in a dataset where articles are nested into outlets, which in turn are nested into topics. Recalling the factors of the bias subcategory cognitive bias. One of the factors, level of involvement, states that the more an individual is involved with a topic, the more likely it is that news is perceived as biased. Therefore, one logical assumption is that articles on generally more polarizing topics receive more comments which have stronger pronounced characteristics.

Furthermore, the attribute scores requested from the Perspective API have not been further examined in this thesis due to limited time. However, they might provide valuable insights for future research as they allow a more fine-graded differentiation of comment characteristics. Considering the definitions of the attributes (cf. Table A.1), some of them can clearly be identified as types of hateful language, for example, "toxicity", "severe toxicity", "identity attack", "insult", "profanity", or "threat". This approach ties in, for example, with the existing work of Davidson, et al. [33], who differentiate between hateful language and offensive language.

In general, with regard to future work, considering additional information sources might reveal valuable insights. For example, approaches exists where researchers evaluate the bias of an article by classifying the article's headline [106]. One could extend this approach, for example, by examining whether the article's headline influences how strongly the comment characteristics are pronounced, for example, by observing whether more luridly framed headlines attract more emotionally charged comments.

Lastly, considering the fact that Twitter is an enormously large database that collects all kinds of user-generated data, conducting a user-group analysis might provide additional valuable insights. Adapting ideas of already existing user group analysis approaches [75, 105], one could investigate demographics of users like the user's age, gender, or political orientation. This user-specific information can then be used in different contexts, for example, to investigate within- and between-group dynamics of the comments on articles of different outlets.

6 CONCLUSION

Since the Internet has become increasingly important to our society, the media environment has changed. A lot of information exchange takes place online, be it in the private sphere to stay connected with friends and family, but also in the professional sphere, for example, companies that increasingly hold meetings online. This development has even been reinforced due to the ongoing COVID-19 pandemic, and online communication channels have become even more deeply embedded in our society. Social media platforms like Facebook or Twitter are widespread mediums for all kinds of information gathering, with increasing popularity for news consumption. It is, therefore, no surprise that news outlets are increasingly relying on online media channels to disseminate their articles. Consequently, the amount of news available online is greater and more diverse than ever before.

The downside of these online channels is that anyone can write news stories and share it with a large potential readership - with almost no control. This uncontrolled flow of information is problematic, as it leads to the increased sharing of false or inaccurate information. The ongoing COVID-19 pandemic has once again made it frighteningly clear how quickly false information can spread and how dangerous such misinformation can be. However, fake news is just one of many ways how biased news can manifest. Nonetheless, the consequences media bias has on our society can be severe. Hence, understanding the underlying structures behind media bias becomes increasingly important.

This thesis pursues exactly this goal by making several contributions to the media bias research. On the one hand, with the help of the presented theoretical framework, the categorization of different bias types is simplified. It facilitates understanding how different bias types are related to each other and how they can be distinguished. Furthermore, with the help of a newly created media bias dataset, it is examined whether comment characteristics can be an indicator of an article's bias. The results provide evidence that the level of hate in the comments is associated with more biased articles. Specifically, talking about the political bias of an article as well as its reliability. The results also show that this hate-bias relationship is reinforced by the news outlet's stance. The thesis provides evidence that outlets that generally have a more pronounced political bias and outlets that are considered reliable show a stronger expression of this hate-bias relationship.

In conclusion, this work contributes to making news coverage more transparent. Working towards a media environment that encourages fair and neutral reporting becomes increasingly important the more our lives shift towards the online world.

A APPENDIX

A.1 ADDITIONAL TABLES

A.1.1 *Description of the Perspective API Attributes*

In the following Table A.1 a detailed definition of each Perspective API attribute is provided.

Table A.1: The Perspective API Attributes

Attribute	**Description**
Toxicity	"A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
Severe Toxicity	"A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words."
Identity Attack	"Negative or hateful comments targeting someone because of their identity."
Insult	"Insulting, inflammatory, or negative comment towards a person or a group of people."
Profanity	"Swear words, curse words, or other obscene or profane language."
Threat	"Describes an intention to inflict pain, injury, or violence against an individual or group."
Sexually Explicit*	"Contains references to sexual acts, body parts, or other lewd content."

Continued on next page

Table A.1: The Perspective API Attributes (cont.)

Attribute	Description
Flirtation*	"Pickup lines, complimenting appearance, subtle sexual innuendos, etc."
Attack on Author**	"Attack on the author of an article or post."
Attack on Commenter**	"Attack on fellow commenter."
Incoherent**	"Difficult to understand, nonsensical."
Inflammatory**	"Intending to provoke or inflame."
Likely to Reject**	"Overall measure of the likelihood for the comment to be rejected according to the NYT's moderation."
Obscene**	"Obscene or vulgar language such as cursing."
Spam**	"Irrelevant and unsolicited commercial content."
Unsubstantial**	"Trivial or short comments"
Note:	
*	Attributes marked with * are experimental attributes "that have not been tested as thoroughly as production attributes" [79].
**	Attributes marked with ** have only been trained on New York Times (NYT) data. Hence, their functionality might be not that broad [79].

Note: This Table contains the descriptions of all 16 Perspective API attributes. All descriptions are cited from the Perspective's website [79].

A.1.2 *All Rated News Outlets*

The following Table A.2 provides a listing of all news outlets that have been rated by ad fontes media. However, not all outlets are included in the final dataset. See the Table's note for more information.

Table A.2: Overview of all Rated Outlets

	Outlet	
19th News	Fortune	RedState
ABC News	Forward	Reuters
Advocate Magazine	Fox Business	Right Wing Watch
Agence France-Presse	Fox News	Roll Call
Al Jazeera	Glamour	RT
AL.com	Glenn Beck**	Salon
AlterNet	Glenn Greenwald	Salt Lake Tribune***
American Greatness	Global News	San Diego Union-Tribune
American Independent	Good News Network	San Francisco Chronicle
American Prospect	Harper's Bazaar	SeattlePI
American Thinker**	Hartford Courant	Second Nexus***
AP	Hawaii News Now	SF Examiner
Arizona Daily Star	Heavy	SFGate
Army Times	High Country News	Shadowproof
ARS Technica	Hill Reporter	Sky News
Aspen Times	Houston Chronicle	Slate
Atlanta Black Star	HuffPost	Smithsonian Magazine
Atlanta Journal-Constitution	In These Times	Snopes
Axios	Independent Journal Review***	Sojourners
AZ Central	Indianapolis Star	South China Morning Post
Baltimore Sun	Inquisitr	Sputnik International News
BBC	Inside Climate News	St. Louis Post-Dispatch
Bearing Arms	Insider	Star Tribune-Minneapolis
Before It's News	Jacobin	Stars and Stripes
Big League Politics	Jezebel	Sun Sentinel

Continued on next page

Table A.2: Overview of all Rated Outlets (cont.)

	Outlet	
Bill O'Reilly	Judicial Watch	Syracuse Post-Standard
Billboard	Just the News	Talking Points Memo
Bipartisan Report**	Kansas City Star	Tampa Bay Times
Bloomberg Government**	LA Times	Tangle
Bloomberg News	LA Weekly	TechCrunch
Boing Boing*	Laconia Daily Sun***	Teen Vogue
Boston Globe	Las Vegas Review-Journal	Tennessean
Boston Herald	Liberty Nation	The American Conservative
Breitbart	Life News	The American Spectator
Bring Me the News	LifeZette	The Atlantic
BuzzFeed News	MarketWatch	The Bulwark
Capitol Weekly	MEDIAite	The Business Journals***
Cato Institute	Meidas Touch	The Christian Post
CBN	Mercury News	The College Fix
CBS News	Mic	The Dispatch
CFO	Military Times	The Economist
Charleston Gazette-Mail	Milwaukee Journal Sentinel	The Evening Times*
Chicago Sun-Times	Montana Free Press	The Federalist
Chicago Tribune	Mother Jones	The Guardian
Chicks on the Right	MSNBC	The Hill
Christian Science Monitor***	National Catholic Register	The Independent
Christianity Today	National File	The Intercept
Chron.com***	National Review	The Liberty Loft***
Cleveland.com	NBC News	The Nation
CNBC	New Republic	The New American**
CNET	New Statesman	The New York Times
CNN**	New York Daily News	The New Yorker
CNSNews	New York Magazine***	The Oregonian
Colorado Daily**	New York Post	The Progressive

Continued on next page

Table A.2: Overview of all Rated Outlets (cont.)

	Outlet	
Columbia Journalism Review	New York Sun*	The Right Scoop
Comic Sands*	News and Guts	The Root
Common Dreams	NewsBusters	The Skimm
Conservative Review	Newsday***	The Stream
Consortium News	Newser***	The Trace
Cosmopolitan	Newsmax	The Verge
CounterPunch	NewsNation Now	The Village Voice
Crooks and Liars	NewsOne	The Weather Channel
Current Affairs	NewsPunch**	The Week
Daily Beast	Newsweek	TheGrio
Daily Caller	Newsy	Time Magazine
Daily Dot	NJ.com	TMZ
Daily Herald-Chicago***	NOLA.com	Townhall
Daily Kos	NowThis News	Truthout
Daily Mail**	NPR	Turning Point USA**
Daily Signal	OAN Network	Twitchy
Daily Torch	Occupy Democrats	UPI
Daily Wire	Omaha World-Herald	Upworthy
Dallas Morning News	Orlando Sentinel	US News and World Report
Deadline	OZY	USA Today
Defense News	Palmer Report	Vanity Fair
Democracy Now	Patch***	Variety
Denver Post	PBS	Vice
Deseret News	Pittsburgh Post-Gazette	Vogue
Detroit Free Press	PJ Media	Voice of America
Detroit News	Politico	Vox
Education Week	Politicus USA	Wall Street Journal
Elite Daily	Politifact	Washington Blade
Elle	Popsugar	Washington Examiner

Continued on next page

Table A.2: Overview of all Rated Outlets (cont.)

Outlet
Engadget
Epoch Times
Esquire
FAIR
Fair Observer
Financial Buzz
Financial Times
Fiscal Times**
FiveThirtyEight
Forbes
Foreign Affairs
Foreign Policy
Popular Information
Poynter
PragerU
ProPublica
Quartz
Quillette
Radio Times
Rasmussen Reports
Raw Story
Real News Network
RealClear Politics
Reason
Washington Free Beacon
Washington Monthly
Washington Post
Washington Times
Washingtonian
Western Journal
WIRED
WND
Wonkette
ZeroHedge

Note: This Table lists all news outlets that have been rated by ad fontes media.
* No tweets have been posted within the dedicated time frame.
** No original tweets have been found.
*** No comments or quoted retweets have been found for all articles of this outlet.

A.1.3 *The Attributes of the Final Dataset*

In Section 4.3, the final dataset created throughout this work has been presented. The following Table A.3 describes each attribute of the dataset. For each numerical attribute, its value range is specified. The definitions of the Perspective API attribute are stated in Table A.1, hence it has been renounced to define them again in this Table.

Table A.3: Definition of all Parameters of the Final Dataset

Parameter	**Description**	**Value Range**
id	The unique tweet id of the comment.	
text	The text of the comment.	
tweet_id	The id of the original tweet to which the comment belongs.	
title	The headline of the article which is referenced by the original tweet.	
outlet	The news outlet which published the article and posted the original tweet.	
twitter_handle	The Twitter user name of the news outlet	
article_url	The URL of the article which is referenced by the original tweet.	
adfontes_url	The URL to the outlet's subpage on ad fontes media's website.	
bias_score	The bias score of the article which is referenced by the original tweet. This score indicates the political bias. A score of 0 indicates no bias.	$[-42, 42]$
reliability_score	The reliability score of the article which is referenced by the original tweet. This score indicates the truthfulness.	$[0, 64]$
pos_score_hate	The positive hate score of the tweet returned by the XLNet-based classifier. The closer the value to 1, the higher the likelihood that the respective tweet is hate.	$[0, 1]$
neg_score_hate	The negative hate score of the tweet returned by the XLNet-based classifier. The closer the value to 1, the higher the likelihood that the respective tweet is non-hate.	$[0, 1]$

Continued on next page

Table A.3: Definition of all Parameters of the Final Dataset (cont.)

Parameter	Description	Value Range
hate_value	The label indicating whether the tweet is hate or non-hate.	(non-hate, hate)
pos_score_sentiment	The positive sentiment score of the tweet returned by the XLNet-based classifier. The closer the value to 1, the higher the likelihood that the respective tweet is positive.	$[0, 1]$
neg_score_sentiment	The negative sentiment score of the tweet returned by the XLNet-based classifier. The closer the value to 1, the higher the likelihood that the respective tweet is negative.	$[0, 1]$
sentiment	The label indicating whether the tweet is positive or negative	(negative, positive)
INSULT	The score for the Perspective API attribute Insult*	$[0, 1]$
LIKELY_TO_REJECT	The score for the Perspective API attribute Likely to Reject*	$[0, 1]$
IDENTITY_ATTACK	The score for the Perspective API attribute Identity Attack*	$[0, 1]$
SEVERE_TOXICITY	The score for the Perspective API attribute Severe Toxicity*	$[0, 1]$
THREAT	The score for the Perspective API attribute Threat*	$[0, 1]$
FLIRTATION	The score for the Perspective API attribute Flirtation*	$[0, 1]$
TOXICITY	The score for the Perspective API attribute Toxicity*	$[0, 1]$
ATTACK_ON_COMMENTER	The score for the Perspective API attribute Attack on Commenter*	$[0, 1]$
SEXUALLY_EXPLICIT	The score for the Perspective API attribute Sexually Explicit*	$[0, 1]$
SPAM	The score for the Perspective API attribute Spam*	$[0, 1]$
INCOHERENT	The score for the Perspective API attribute Incoherent*	$[0, 1]$

Continued on next page

Table A.3: Definition of all Parameters of the Final Dataset (cont.)

Parameter	Description	Value Range
UNSUBSTANTIAL	The score for the Perspective API attribute Unsubstantial*	$[0,1]$
ATTACK_ON_AUTHOR	The score for the Perspective API attribute Attack on Author*	$[0,1]$
PROFANITY	The score for the Perspective API attribute Profanity*	$[0,1]$
INFLAMMATORY	The score for the Perspective API attribute Inflammatory*	$[0,1]$
OBSCENE	The score for the Perspective API attribute Obscene*	$[0,1]$
bias_class	The bias class of the outlet.	**
reliability_class	The reliability class of the outlet.	***
overall_reliability	The overall reliability score of the outlet.	$[0,64]$
overall_bias	The overall bias score of the outlet	$[-42,42]$
polarity_strength	The polarity strength of the tweet. The closer the value to 1, the stronger the polarity of the tweet (positive or negative).	$[0,1]$

Note: This Table provides an explanation for all attributes of the final dataset.

* Please refer to Table A.1 for detailed descriptions of the Perspective API attributes.

* The bias classes for outlets are: 1) most extreme left, 2) hyper-partisan left, 3) skews left, 4) middle or balanced bias, 5) skews right, 6) hyper-partisan right, 7) most extreme right

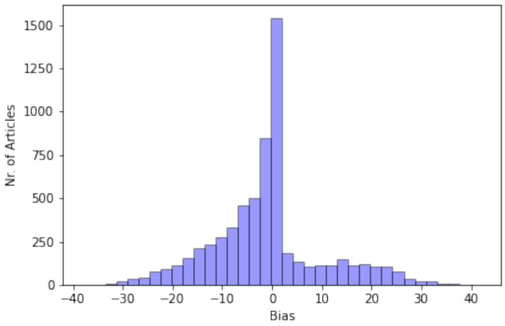
** The reliability classes for outlets are: 1) original fact reporting, 2) fact reporting, 3) complex analysis or mix of fact reporting and analysis, 4) analysis or high variation in reliability, 5) opinion or high variation in reliability, 6) selective or incomplete story/unfair persuasion/propaganda, 7) contains misleading information, 8) contains inaccurate/fabricated information

A.2 ADDITIONAL FIGURES

A.2.1 *Distribution Plots for the Initial Dataset*

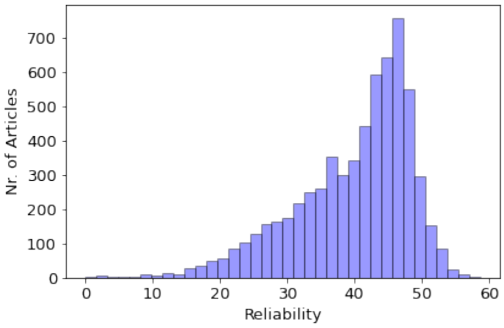
The following plots refer to the initial dataset that have been collected from ad fontes media's website (cf. Section 4.1.1), before further excluding articles and outlets due to the Twitter data collection process. The figures show the distribution of bias and reliability scores among all articles (Figure A.1 and Figure A.2), and among all outlets (Figure A.3 and Figure A.4). The initial dataset contains 6,345 articles of 283 news outlets.

Figure A.1: Distribution of Bias Scores among all Articles



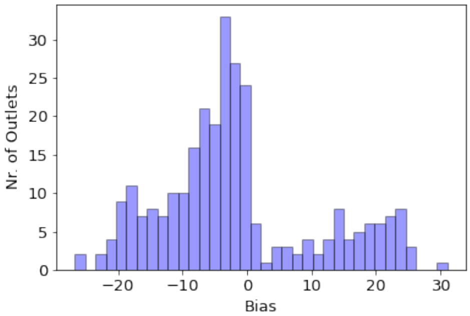
Note: This Figure shows the distribution of the article bias scores over the dataset, where the bias ranges from -42 (hyperpartisan left) to +42 (hyperpartisan right).

Figure A.2: Distribution of Reliability Scores among all Articles



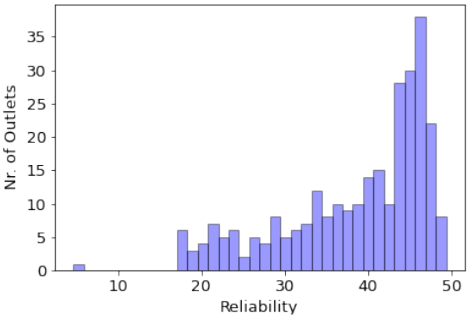
Note: This Figure shows the distribution of the article reliability scores over the dataset, where the score ranges from 0 (most unreliable) to +64 (most reliable).

Figure A.3: Distribution of Bias Scores among all Outlets



Note: This Figure shows the distribution of the outlet bias scores over the dataset, where the bias ranges from -42 (hyperpartisan left) to +42 (hyperpartisan right).

Figure A.4: Distribution of Reliability Scores among all Outlets

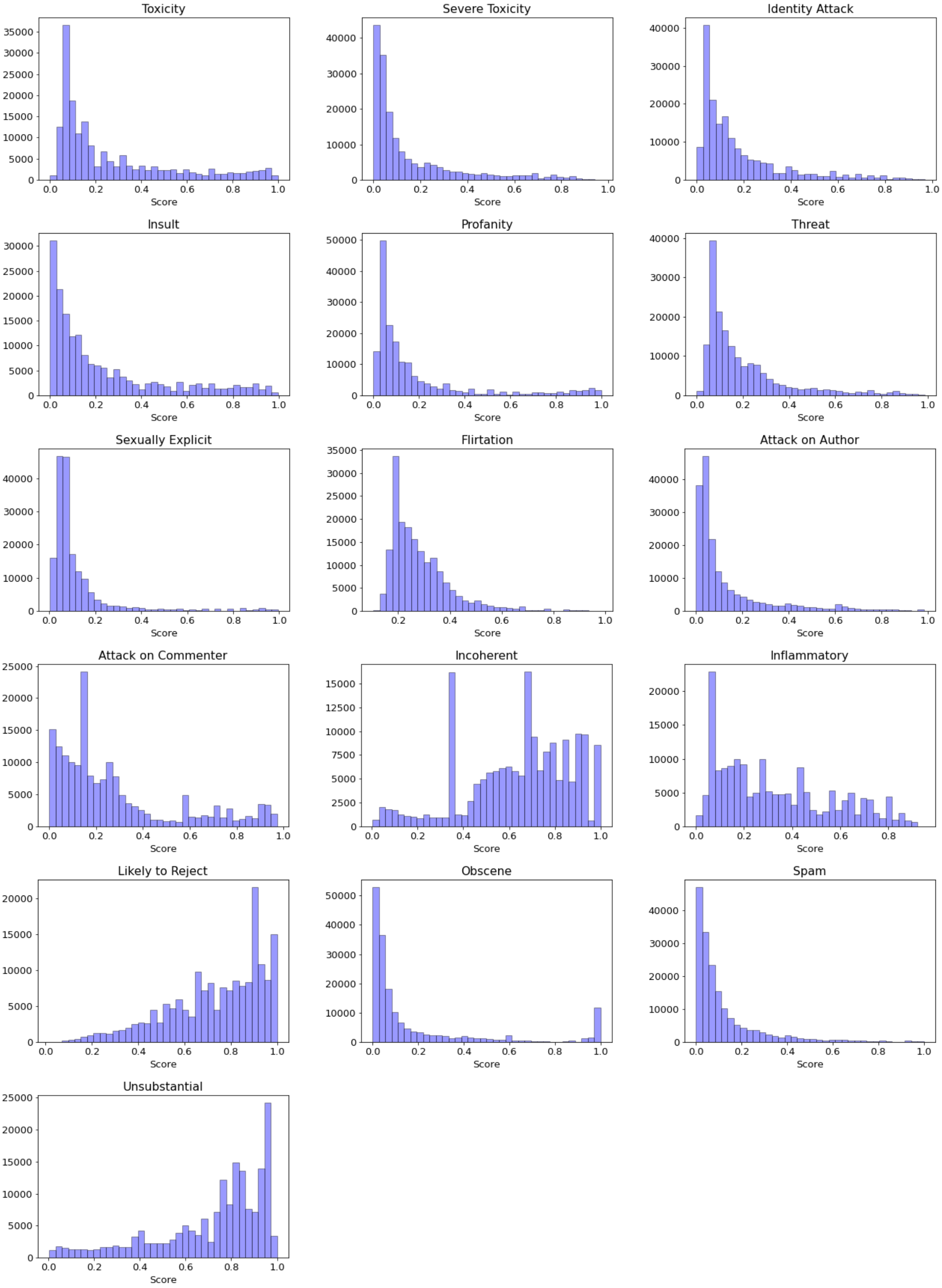


Note: This Figure shows the distribution of the outlet reliability scores over the dataset, where the score ranges from 0 (most unreliable) to +64 (most reliable).

A.2.2 *Distribution Plots for the Perspective API Attributes*

In the following, the distribution plots for all 16 Perspective API attribute scores are shown. This Figure shows the distributions of all 16 Perspective API attribute scores among all tweets. The scores indicate how likely a tweet is perceived as said attribute, where 0 indicates a likelihood of 0% and 1 indicates a likelihood of 100%.

Figure A.5: Distributions of the 16 Perspective API Attribute Scores

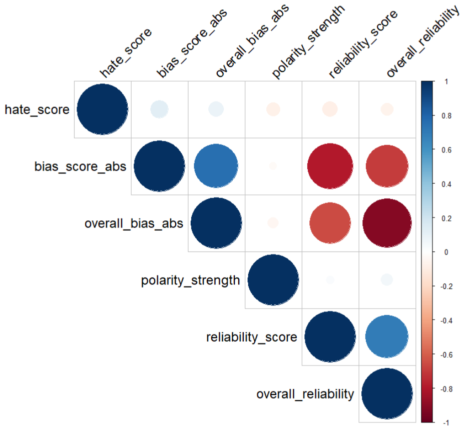


Note: This Figure shows the distributions of all 16 Perspective API attribute scores among all tweets. The scores indicate how likely a tweet is perceived as said attribute, where 0 indicates a likelihood of 0% and 1 indicates a likelihood of 100%.

A.2.3 *Correlation Matrix: Selected Attributes*

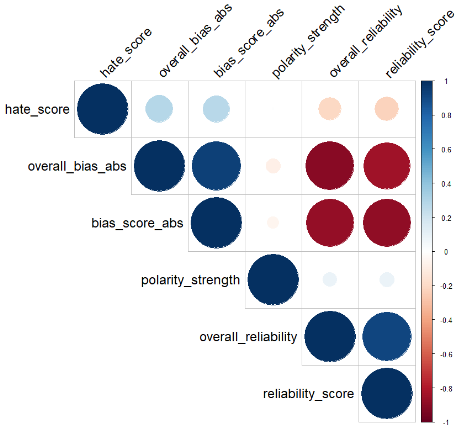
In the following, two correlation matrices for a selected subset of the attributes are shown. In Figure A.6, the correlation matrix grouped on article level is shown. In Figure A.7, the correlation matrix grouped on outlet level is sown. The plots show that correlations exist between individual attributes. However, these are stronger on outlet level.

Figure A.6: Correlation Matrix on Article-Level: Selected Attributes



Note: This Figure shows the correlation matrix for a selected subset of the attributes, grouped at article-level.

Figure A.7: Correlation Matrix on Outlet-Level: Selected Attributes



Note: This Figure shows the correlation matrix for a selected subset of the attributes, grouped at outlet-level.

BIBLIOGRAPHY

[1] Ayush Agarwal, Ashima Yadav, and Dinesh Kumar Vishwakarma. "Multimodal Sentiment Analysis via RNN variants." In: *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science Engineering (BCD)*. 2019, pp. 19–23. DOI: 10.1109/BCD.2019.8885108.

[2] Muhammad Z. Ali, Ehsan-Ul-Haq, Sahar Rauf, Kashif Javed, and Sarmad Hussain. "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis." In: *IEEE Access* 9 (2021), pp. 84296–84305. DOI: 10.1109/ACCESS.2021.3087827.

[3] Omar Ali, Ilias N. Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. "Automating News Content Analysis: An Application to Gender Bias and Readability." In: *Proceedings of the First Workshop on Applications of Pattern Analysis, WAPA 2010, Cumberland Lodge, Windsor, UK, September 1-3, 2010*. Ed. by Tom Diethe, Nello Cristianini, and John Shawe-Taylor. Vol. 11. JMLR Proceedings. JMLR.org, 2010, pp. 36–43. URL: http://proceedings.mlr.press/v11/ali10a.html.

[4] Ahlam Alrehili. "Automatic Hate Speech Detection on Social Media: A Brief Survey." In: *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. 2019, pp. 1–6. DOI: 10.1109/AICCSA47632.2019.9035228.

[5] Karel Jan Alsem, Steven Brakman, Lex Hoogduin, and Gerard Kuper. "The impact of newspapers on consumer confidence: does spin bias exist?" In: *Applied Economics* 40.5 (2008), pp. 531–539. DOI: 10.1080/00036840600707100. eprint: https://doi.org/10.1080/00036840600707100. URL: https://doi.org/10.1080/00036840600707100.

[6] Amarina Ariyanto, Matthew J. Hornsey, and Cindy Gallois. "Group Allegiances and Perceptions of Media Bias: Taking Into Account Both the Perceiver and the Source." In: *Group Processes & Intergroup Relations* 10.2 (2007), pp. 266–279. DOI: 10.1177/1368430207074733. eprint: https://doi.org/10.1177/1368430207074733. URL: https://doi.org/10.1177/1368430207074733.

[7] Dorothee Arlt, Caroline Dalmus, and Julia Metag. "Direct and Indirect Effects of Involvement on Hostile Media Perceptions in the Context of the Refugee Crisis in Germany and Switzerland." In: *Mass Communication and Society* 22.2 (2019), pp. 171–195. DOI: 10.1080/15205436.2018.1536791. eprint: https://doi.org/10.1080/15205436.2018.1536791. URL: https://doi.org/10.1080/15205436.2018.1536791.

[8] Laura M. Arpan and Arthur A. Raney. "An Experimental Investigation of News Source and the Hostile Media Effect." In: *Journalism & Mass Communication Quarterly* 80.2 (2003), pp. 265–281. DOI: 10.1177/107769900308000203. eprint: https://doi.org/10.1177/107769900308000203. URL: https://doi.org/10.1177/107769900308000203.

[9] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. "Exposure to opposing views on social media can increase political polarization." In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221. DOI: 10.1073/pnas.1804840115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1804840115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1804840115.

[10] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis." In: *Future Generation Computer Systems* 115 (2021), pp. 279–294. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.08.005. URL: https://www.sciencedirect.com/science/article/pii/S0167739X20309195.

[11] Matthew A. Baum and Phil Gussin. "In the Eye of the Beholder: How Information Shortcuts Shape Individual Perceptions of Bias in the Media." In: *Quarterly Journal of Political Science* (2008). URL: https://www.nowpublishers.com/article/Details/QJPS-7010.

[12] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. "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*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 1472–1482. DOI: 10.3115/v1/N15-1171. URL: https://aclanthology.org/N15-1171.

[13] Dan Bernhardt, Stefan Krasa, and Mattias Polborn. "Political polarization and the electoral effects of media bias." In: *Journal of Public Economics* 92.5 (2008), pp. 1092–1104. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2008.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0047272708000236.

[14] Michael R. Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn, and Rosaria Silipo. *Guide to Intelligent Data Science*. 2nd ed. Springer, Cham, 2020, pp. XIII–420. DOI: https://doi.org/10.1007/978-3-030-45574-3.

[15] Camiel J. Beukeboom and Christian Burgers. *Linguistic Bias*. July 2017. DOI: 10.1093/acrefore/9780190228613.013.439. URL: https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-439.

[16] Chetashri Bhadane, Hardi Dalal, and Heenal Doshi. "Sentiment Analysis: Measuring Opinions." In: *Procedia Computer Science* 45 (2015). International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 808–814. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.03.159. URL: https://www.sciencedirect.com/science/article/pii/S1877050915003956.

[17] Jordan J. Bird, Anikó Ekárt, and Diego R. Faria. "Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification." In: *Journal of Ambient Intelligence and Humanized Computing* (2021). DOI: https://doi.org/10.1007/s12652-021-03439-8.

[18] Fernando Blanco. "Cognitive Bias." In: *Encyclopedia of Animal Cognition and Behavior*. Ed. by Jennifer Vonk and Todd Shackelford. Cham: Springer International Publishing, 2017, pp. 1–7. ISBN: 978-3-319-47829-6. DOI: 10.1007/978-3-319-47829-6_1244-1. URL: https://doi.org/10.1007/978-3-319-47829-6_1244-1.

[19] Ceren Budak, Sharad Goel, and Justin M. Rao. "Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis." In: *Public Opinion Quarterly* 80.S1 (Apr. 2016), pp. 250–271. ISSN: 0033-362X. DOI: 10.1093/poq/nfw007. URL: https://doi.org/10.1093/poq/nfw007.

[20] Pete Burnap, Omer F. Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. "Detecting tension in online communities with computational Twitter analysis." In: *Technological Forecasting and Social Change* 95 (2015), pp. 96–108. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2013.04.013. URL: https://www.sciencedirect.com/science/article/pii/S0040162513000899.

[21] Dustin P. Calvillo, Abraham M. Rutchick, and Ryan J. B. Garcia. "Individual Differences in Belief in Fake News about Election Fraud after the 2020 U.S. Election." In: *Behavioral Sciences* 11.12 (2021). ISSN: 2076-328X. DOI: 10.3390/bs11120175. URL: https://www.mdpi.com/2076-328X/11/12/175.

[22] Neal Caren, Kenneth T. Andrews, and Todd Lu. "Contemporary Social Movements in a Hybrid Media Environment." In: *Annual Review of Sociology* 46.1 (2020), pp. 443–465. DOI: 10.1146/annurev-soc-121919-054627. eprint: https://doi.org/10.1146/annurev-soc-121919-054627. URL: https://doi.org/10.1146/annurev-soc-121919-054627.

[23] Laia Castro, David Nicolas Hopmann, and Lilach Nir. "Whose media are hostile? The spillover effect of interpersonal discussions on media bias perceptions." In: *Communications* 46.4 (2021), pp. 540–563. DOI: doi:10.1515/commun-2019-0140. URL: https://doi.org/10.1515/commun-2019-0140.

[24] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. "Detecting Media Bias in News Articles using Gaussian Bias Distributions." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4290–4300. DOI: 10.18653/v1/2020.findings-emnlp.383. URL: https://aclanthology.org/2020.findings-emnlp.383.

[25] Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. "Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity." In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. Online: Association for Computational Linguistics, Nov. 2020, pp. 149–154. DOI: 10.18653/v1/2020.nlpcss-1.16. URL: https://aclanthology.org/2020.nlpcss-1.16.

[26] Lamogha Chiazor, Geeth de Mel, Graham White, Gwilym Newton, Joe Pavitt, and Richard J. Tomsett. "An Automated Framework to Identify and Eliminate Systemic Racial Bias in the Media." In: *CEUR Workshop Proceedings* 2812 (Feb. 2021), pp. 32–36. ISSN: 1613-0073. URL: http://ceur-ws.org/Vol-2812/.

[27] Marta Costa-jussa. "An analysis of gender bias studies in natural language processing." In: *Nature Machine Intelligence* 1 (Nov. 2019), pp. 495–496. DOI: https://doi.org/10.1038/s42256-019-0105-5.

[28] Andres Cremisini, Daniela Aguilar, and Mark A. Finlayson. "A Challenging Dataset for Bias Detection: The Case of the Crisis in the Ukraine." In: *Social, Cultural, and Behavioral Modeling*. Ed. by Robert Thomson, Halil Bisgin, Christopher Dancy, and Ayaz Hyder. Cham: Springer International Publishing, 2019, pp. 173–183. ISBN: 978-3-030-21741-9.

[29] Dave D'Alessio. "An Experimental Examination of Readers' Perceptions of Media Bias." In: *Journalism & Mass Communication Quarterly* 80.2 (2003), pp. 282–294. DOI: 10.1177/107769900308000204. eprint: https://doi.org/10.1177/107769900308000204. URL: https://doi.org/10.1177/107769900308000204.

[30] Dave D'Alessio and Mike Allen. "Media Bias in Presidential Elections: A Meta-Analysis." In: *Journal of Communication* 50.4 (Jan. 2000), pp. 133–156. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.2000.tb02866.x. eprint: https://academic.oup.com/joc/article-pdf/50/4/133/22335481/jjnlcom0133.pdf. URL: https://doi.org/10.1111/j.1460-2466.2000.tb02866.x.

[31] Jamell Dacon and Haochen Liu. “Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles.” In: *Companion Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 385–392. ISBN: 9781450383134. URL: https://doi.org/10.1145/3442442.3452325.

[32] Russell J. Dalton, Paul A. Beck, and Robert Huckfeldt. “Partisan Cues and the Media: Information Flows in the 1992 Presidential Election.” In: *The American Political Science Review* 92.1 (1998), pp. 111–126. ISSN: 00030554, 15375943. URL: http://www.jstor.org/stable/2585932.

[33] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. “Automated Hate Speech Detection and the Problem of Offensive Language.” In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017), pp. 512–515. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

[34] Junetta Davis. “Sexist Bias in Eight Newspapers.” In: *Journalism Quarterly* 59.3 (1982), pp. 456–460. DOI: 10.1177/107769908205900316. URL: https://doi.org/10.1177/107769908205900316.

[35] Stefano DellaVigna and Ethan Kaplan. *The Fox News Effect: Media Bias and Voting*. Working Paper 12169. National Bureau of Economic Research, Apr. 2006. DOI: 10.3386/w12169. URL: http://www.nber.org/papers/w12169.

[36] *Detecting Insults in Social Commentary*. 2013. URL: https://www.kaggle.com/c/detecting-insults-in-social-commentary.

[37] Travis L. Dixon and Daniel Linz. “Race and the Misrepresentation of Victimization on Local Television News.” In: *Communication Research* 27.5 (2000), pp. 547–573. DOI: 10.1177/009365000027005001. eprint: https://doi.org/10.1177/009365000027005001. URL: https://doi.org/10.1177/009365000027005001.

[38] Marko Dragojevic, Alexander Sink, and Dana Mastro. “Evidence of Linguistic Intergroup Bias in U.S. Print News Coverage of Immigration.” In: *Journal of Language and Social Psychology* 36.4 (2017), pp. 462–472. DOI: 10.1177/0261927X16666884. eprint: https://doi.org/10.1177/0261927X16666884. URL: https://doi.org/10.1177/0261927X16666884.

[39] James N. Druckman and Michael Parkin. “The Impact of Media Bias: How Editorial Slant Affects Voters.” In: *The Journal of Politics* 67.4 (2005), pp. 1030–1049. ISSN: 00223816, 14682508. URL: http://www.jstor.org/stable/10.1111/j.1468-2508.2005.00349.x.

[40] Zulfadzli Drus and Haliyana Khalid. "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review." In: *Procedia Computer Science* 161 (2019). The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia, pp. 707–714. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.11.174. URL: https://www.sciencedirect.com/science/article/pii/S187705091931885X.

[41] Akash Dutt Dubey. *Twitter Sentiment Analysis during COVID-19 Outbreak*. Apr. 2020. DOI: 10.2139/ssrn.3572023. URL: https://ssrn.com/abstract=3572023.

[42] Elizabeth Dubois and Grant Blank. "The echo chamber is overstated: the moderating effect of political interest and diverse media." In: *Information, Communication & Society* 21.5 (2018), pp. 729–745. DOI: 10.1080/1369118X.2018.1428656. eprint: https://doi.org/10.1080/1369118X.2018.1428656. URL: https://doi.org/10.1080/1369118X.2018.1428656.

[43] Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. "One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences." In: *Communication Research* 44.8 (2017), pp. 1125–1148. DOI: 10.1177/0093650215614364. eprint: https://doi.org/10.1177/0093650215614364. URL: https://doi.org/10.1177/0093650215614364.

[44] Craig Enders and Davood Tofighi. "Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at An Old Issue." In: *Psychological Methods* 12.2 (2007), pp. 121–138. DOI: 10.1037/1082-989X.12.2.121.

[45] Kenneth C. Enevoldsen and Lasse Hansen. "Analysing Political Biases in Danish Newspapers Using Sentiment Analysis." In: *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift* 2.2 (July 2017), pp. 87–98. URL: https://tidsskrift.dk/lwo/article/view/96014.

[46] Robert M. Entman. "Framing Bias: Media in the Distribution of Power." In: *Journal of Communication* 57.1 (Feb. 2007), pp. 163–173. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.2006.00336.x. eprint: https://academic.oup.com/joc/article-pdf/57/1/163/22326478/jjnlcom0163.pdf. URL: https://doi.org/10.1111/j.1460-2466.2006.00336.x.

[47] William P. Eveland Jr. and Dhavan V. Shah. "The Impact of Individual and Interpersonal Factors on Perceived News Media Bias." In: *Political Psychology* 24.1 (2003), pp. 101–117. DOI: https://doi.org/10.1111/0162-895X.00318. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/0162-895X.00318.

[48] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. "In Plain Sight: Media Bias Through the Lens of Factual Reporting." In: *CoRR* abs/1909.02670 (2019). arXiv: 1909.02670. URL: http://arxiv.org/abs/1909.02670.

[49] Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. "A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias." In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 3007–3014. ISBN: 9781450368599. DOI: 10.1145/3340531.3412876. URL: https://doi.org/10.1145/3340531.3412876.

[50] Johan Galtung and Mari Holmboe Ruge. "The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers." In: *Journal of Peace Research* 2.1 (1965), pp. 64–90. DOI: 10.1177/002234336500200104. eprint: https://doi.org/10.1177/002234336500200104. URL: https://doi.org/10.1177/002234336500200104.

[51] William A. Gamson and Andre Modigliani. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." In: *American Journal of Sociology* 95.1 (1989), pp. 1–37. DOI: 10.1086/229213. eprint: https://doi.org/10.1086/229213. URL: https://doi.org/10.1086/229213.

[52] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. "Target-Dependent Sentiment Classification With BERT." In: *IEEE Access* 7 (2019), pp. 154290–154299. DOI: 10.1109/ACCESS.2019.2946594.

[53] Alan S. Gerber, Dean Karlan, and Daniel Bergan. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." In: *American Economic Journal: Applied Economics* 1.2 (Apr. 2009), pp. 35–52. DOI: 10.1257/app.1.2.35. URL: https://www.aeaweb.org/articles?id=10.1257/app.1.2.35.

[54] Sarah Gershon. "When Race, Gender, and the Media Intersect: Campaign News Coverage of Minority Congresswomen." In: *Journal of Women, Politics & Policy* 33.2 (2012), pp. 105–125. DOI: 10.1080/1554477X.2012.667743. eprint: https://doi.org/10.1080/1554477X.2012.667743. URL: https://doi.org/10.1080/1554477X.2012.667743.

[55] Ona de Gibert, Naiara Pérez, Aitor García Pablos, and Montse Cuadros. "Hate Speech Dataset from a White Supremacy Forum." In: *CoRR* abs/1809.04444 (2018). arXiv: 1809.04444. URL: http://arxiv.org/abs/1809.04444.

[56] Dennis Njagi Gitari, Zhang Zuping, Damien Hanyurwimfura, and Jun Long. "A Lexicon-based Approach for Hate Speech Detection." In: *International Journal of Multimedia and Ubiquitous Engineering* 10.4 (Apr. 2015), pp. 215–230. DOI: 10.14257/ijmue.2015.10.4.21. URL: http://dx.doi.org/10.14257/ijmue.2015.10.4.21.

[57] Carroll J. Glynn and Michael E. Huge. “How Pervasive Are Perceptions of Bias? Exploring Judgments of Media Bias in Financial News.” In: *International Journal of Public Opinion Research* 26.4 (Feb. 2014), pp. 543–553. ISSN: 0954-2892. DOI: 10.1093/ijpor/edu004. eprint: https://academic.oup.com/ijpor/article-pdf/26/4/543/2191013/edu004.pdf. URL: https://doi.org/10.1093/ijpor/edu004.

[58] Alec Go, Richa Bhayani, and Lei Huang. “Twitter sentiment classification using distant supervision.” In: (2009), pp. 1–6.

[59] Stephan Greene and Philip Resnik. “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*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 503–511. URL: https://aclanthology.org/N09-1057.

[60] Albert C. Gunther, Cindy T. Christen, Janice L. Liebhart, and Stella Chih-Yun Chia. “Congenial Public, Contrary Press, and Biased Estimates of the Climate of Opinion.” In: *The Public Opinion Quarterly* 65.3 (2001), pp. 295–320. ISSN: 0033362X, 15375331. URL: http://www.jstor.org/stable/3078822.

[61] Albert C. Gunther and Janice L. Liebhart. “Broad Reach or Biased Source? Decomposing the Hostile Media Effect.” In: *Journal of Communication* 56.3 (Aug. 2006), pp. 449–466. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.2006.00295.x. eprint: https://academic.oup.com/joc/article-pdf/56/3/449/22325694/jjnlcom0449.pdf. URL: https://doi.org/10.1111/j.1460-2466.2006.00295.x.

[62] Albert C. Gunther, Bryan McLaughlin, Melissa R. Gotlieb, and David Wise. “Who Says What to Whom: Content Versus Source in the Hostile Media Effect.” In: *International Journal of Public Opinion Research* 29.3 (May 2016), pp. 363–383. ISSN: 0954-2892. DOI: 10.1093/ijpor/edw009. eprint: https://academic.oup.com/ijpor/article-pdf/29/3/363/19649789/edw009.pdf. URL: https://doi.org/10.1093/ijpor/edw009.

[63] Albert C. Gunther and Kathleen Schmitt. “Mapping Boundaries of the Hostile Media Effect.” In: *Journal of Communication* 54.1 (Jan. 2006), pp. 55–70. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.2004.tb02613.x. eprint: https://academic.oup.com/joc/article-pdf/54/1/55/22330555/jjnlcom0055.pdf. URL: https://doi.org/10.1111/j.1460-2466.2004.tb02613.x.

[64] Felix Hamborg and Karsten Donnay. “NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles.” In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics,

Apr. 2021, pp. 1663–1675. DOI: 10.18653/v1/2021.eacl-main.142. URL: https://aclanthology.org/2021.eacl-main.142.

[65] Felix Hamborg, Karsten Donnay, and Bela Gipp. "Automated identification of media bias in news articles: an interdisciplinary literature review." In: *International Journal on Digital Libraries* (2019), pp. 391–415. DOI: 10.1007/s00799-018-0261-y. URL: https://doi.org/10.1007/s00799-018-0261-y.

[66] Felix Hamborg, Norman Meuschke, and Bela Gipp. "Matrix-Based News Aggregation: Exploring Different News Perspectives." In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2017, pp. 1–10. DOI: 10.1109/JCDL.2017.7991561.

[67] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. "Automated Identification of Media Bias by Word Choice and Labeling in News Articles." In: *Proceedings of the 18th Joint Conference on Digital Libraries*. JCDL '19. Champaign, Illinois: IEEE Press, 2019, pp. 196–205. ISBN: 9781728115474. DOI: 10.1109/JCDL.2019.00036. URL: https://doi.org/10.1109/JCDL.2019.00036.

[68] I Hemalatha, G. P. Saradhi Varma, and A. Govardhan. "Sentiment Analysis Tool using Machine Learning Algorithms." In: *Computer Science and Engineering* 58 (2013), pp. 14791–14794. ISSN: 2229-712X.

[69] Shirley S. Ho, Andrew R. Binder, Amy B. Becker, Patricia Moy, Dietram A. Scheufele, Dominique Brossard, and Albert C. Gunther. "The Role of Perceptions of Media Bias in General and Issue-Specific Political Participation." In: *Mass Communication and Society* 14.3 (2011), pp. 343–374. DOI: 10.1080/15205436.2010.491933. eprint: https://doi.org/10.1080/15205436.2010.491933. URL: https://doi.org/10.1080/15205436.2010.491933.

[70] J. Brian Houston, Glenn J. Hansen, and Gwendelyn S. Nisbett. "Influence of User Comments on Perceptions of Media Bias and Third-Person Effect in Online News." In: *Electronic News* 5.2 (2011), pp. 79–92. DOI: 10.1177/1931243111407618. eprint: https://doi.org/10.1177/1931243111407618. URL: https://doi.org/10.1177/1931243111407618.

[71] Joop J. Hox. *Multilevel analysis: Techniques and applications*. 2nd ed. New York, NY: Routledge, 2010, pp. 1–392. ISBN: 978–1–84872–845–5.

[72] Christoph Hube and Besnik Fetahu. "Detecting Biased Statements in Wikipedia." In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 1779–1786. ISBN: 9781450356404. DOI: 10.1145/3184558.3191640. URL: https://doi.org/10.1145/3184558.3191640.

[73] Christoph Hube and Besnik Fetahu. "Neural Based Statement Classification for Biased Language." In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Jan. 2019). DOI: 10.1145/3289600.3291018. URL: http://dx.doi.org/10.1145/3289600.3291018.

[74] C. J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 2014), pp. 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[75] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. "Twitter user profiling based on text and community mining for market analysis." In: *Knowledge-Based Systems* 51 (2013), pp. 35–47. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2013.06.020. URL: https://www.sciencedirect.com/science/article/pii/S0950705113002025.

[76] ad fontes media Inc. *ad fontes media*. 2015–2022. URL: https://adfontesmedia.com/ (visited on 10/26/2021).

[77] Ming Jiang, Junlei Wu, Xiangrong Shi, and Min Zhang. "Transformer Based Memory Network for Sentiment Analysis of Web Comments." In: *IEEE Access* 7 (2019), pp. 179942–179953. DOI: 10.1109/ACCESS.2019.2957192.

[78] Zhao Jianqiang and Gui Xiaolin. "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis." In: *IEEE Access* 5 (2017), pp. 2870–2879. DOI: 10.1109/ACCESS.2017.2672677.

[79] Jigsaw. *Perspective*. 2021. URL: https://www.perspectiveapi.com/ (visited on 03/06/2022).

[80] Brandon Joyce and Jing Deng. "Sentiment analysis of tweets for the 2016 US presidential election." In: *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*. 2017, pp. 1–4. DOI: 10.1109/URTC.2017.8284176.

[81] Daniel Kahneman and Amos Tversky. "Choices, values, and frames." In: *American Psychologist* 39.4 (1984), pp. 341–350. DOI: 10.1037/0003-066X.39.4.341. URL: https://doi.org/10.1037/0003-066X.39.4.341.

[82] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[83] Steven Kull, Clay Ramsay, and Evan Lewis. "Misperceptions, the Media, and the Iraq War." In: *Political Science Quarterly* 118.4 (2003), pp. 569–598. ISSN: 00323195. URL: http://www.jstor.org/stable/30035697.

[84] Konstantina Lazaridou, Alexander Löser, Maria Mestre, and Felix Naumann. "Discovering Biased News Articles Leveraging Multiple Human Annotations." In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1268–1277. URL: https://www.aclweb.org/anthology/2020.lrec-1.159.

[85] Eun-Ju Lee. "That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias." In: *Journal of Computer-Mediated Communication* 18.1 (2012), pp. 32–45. DOI: https://doi.org/10.1111/j.1083-6101.2012.01597.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1083-6101.2012.01597.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2012.01597.x.

[86] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. "Mitigating Media Bias through Neutral Article Generation." In: *CoRR* abs/2104.00336 (2021). arXiv: 2104.00336. URL: https://arxiv.org/abs/2104.00336.

[87] Tien-Tsung Lee. "The Liberal Media Myth Revisited: An Examination of Factors Influencing Perceptions of Media Bias." In: *Journal of Broadcasting & Electronic Media* 49.1 (2005), pp. 43–64. DOI: 10.1207/s15506878jobem4901_4. eprint: https://doi.org/10.1207/s15506878jobem4901_4. URL: https://doi.org/10.1207/s15506878jobem4901_4.

[88] Sora Lim, Adam Jatowt, Michael FÃ¤rber, and Masatoshi Yoshikawa. "Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing." In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1478–1484. URL: https://www.aclweb.org/anthology/2020.lrec-1.184.

[89] Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. "Understanding characteristics of biased sentences in news articles." In: *CIKM workshops* (2018).

[90] Anne Maass, Daniela Salvi, Luciano Arcuri, and Gun Semin. "Language use in intergroup contexts: The linguistic intergroup bias." In: *Journal of Personality and Social Psychology* 67.6 (1989), pp. 981–993. DOI: https://doi.org/10.1037/0022-3514.57.6.981.

[91] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. "Hate speech detection: Challenges and solutions." In: *PLOS ONE* 14.8 (Aug. 2019), pp. 1–16. DOI: 10.1371/journal.pone.0221152. URL: https://doi.org/10.1371/journal.pone.0221152.

[92] Héctor Martínez Alonso, Amaury Delamaire, and Benoît Sagot. "Annotating omission in statement pairs." In: *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 41–45. DOI: 10.18653/v1/W17-0805. URL: https://aclanthology.org/W17-0805.

[93] Jörg Matthes, Desirée Schmuck, and Christian von Sikorski. "In the Eye of the Beholder: A Case for the Visual Hostile Media Phenomenon." In: *Communication Research* (2021), pp. 1–25. DOI: 10.1177/00936502211018596. eprint: https://doi.org/10.1177/00936502211018596. URL: https://doi.org/10.1177/00936502211018596.

[94] James Meneghello, Nick Thompson, Kevin Lee, Kok Wai Wong, and Bilal Abu-Salih. “Unlocking Social Media and User Generated Content as a Data Source for Knowledge Management.” In: *International Journal of Knowledge Management (IJKM)* 16.1 (), pp. 101–122. DOI: 10.4018/IJKM.2020010105. URL: http://doi.org/10.4018/IJKM.2020010105.

[95] Seong-Jae Min and John C. Feaster. “Missing Children in National News Coverage: Racial and Gender Representations of Missing Children Cases.” In: *Communication Research Reports* 27.3 (2010), pp. 207–216. DOI: 10.1080/08824091003776289. eprint: https://doi.org/10.1080/08824091003776289. URL: https://doi.org/10.1080/08824091003776289.

[96] Mohiyaddeen and Dr. Shifaulla Siddiqui. “Automatic Hate Speech Detection: A Literature Review.” In: *International Journal of Engineering and Management Research* 11.2 (Apr. 2021), pp. 116–121. DOI: 10.31033/ijemr.11.2.16. URL: https://www.ijemr.net/ojs/index.php/ijemr/article/view/110.

[97] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. “Hate speech detection and racial bias mitigation in social media based on BERT model.” In: *PLOS ONE* 15.8 (Aug. 2020), pp. 1–26. DOI: 10.1371/journal.pone.0237861. URL: https://doi.org/10.1371/journal.pone.0237861.

[98] Sendhil Mullainathan and Andrei Shleifer. *Media Bias*. Working Paper 9295. National Bureau of Economic Research, Oct. 2002. DOI: 10.3386/w9295. URL: http://www.nber.org/papers/w9295.

[99] Sendhil Mullainathan and Andrei Shleifer. “The Market for News.” In: *American Economic Review* 95.4 (Sept. 2005), pp. 1031–1053. DOI: 10.1257/0002828054825619. URL: https://www.aeaweb.org/articles?id=10.1257/0002828054825619.

[100] Diana C. Mutz and Paul S. Martin. “Facilitating Communication across Lines of Political Difference: The Role of Mass Media.” In: *American Political Science Review* 95.1 (2001), pp. 97–114. DOI: 10.1017/S0003055401000223.

[101] M S Neethu and R Rajasree. “Sentiment analysis in twitter using machine learning techniques.” In: *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. 2013, pp. 1–5. DOI: 10.1109/ICCCNT.2013.6726818.

[102] Nic Newman, Dr. Richard Fletcher, Dr. Anne Schulz, Dr. Simge Andi, Dr. Craig T. Robertson, and Prof. Rasmus Kleis Nielsen. “Reuters Institute Digital News Report 2021.” In: *Reuters Institute for the Study of Journalism (RIS)* 10 (2021). URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf.

[103] Finn Årup Nielsen. “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.” In: *CoRR* abs/1103.2903 (2011). arXiv: 1103.2903. URL: http://arxiv.org/abs/1103.2903.

[104] Yilang Peng. "Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision." In: *Journal of Communication* 68.5 (Oct. 2018), pp. 920–941. ISSN: 0021-9916. DOI: 10.1093/joc/jqy041. eprint: https://academic.oup.com/joc/article-pdf/68/5/920/26428783/jqy041.pdf. URL: https://doi.org/10.1093/joc/jqy041.

[105] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. "An analysis of the user occupational class through Twitter content." 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)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 1754–1764. DOI: 10.3115/v1/P15-1169. URL: https://aclanthology.org/P15-1169.

[106] Mazhar Iqbal Rana, Shehzad Khalid, and Muhammad Usman Akbar. "News classification based on their headlines: A review." In: *17th IEEE International Multi Topic Conference 2014*. 2014, pp. 211–216. DOI: 10.1109/INMIC.2014.7097339.

[107] Hannah Rashkin, Sameer Singh, and Yejin Choi. "Connotation Frames: A Data-Driven Investigation." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 311–321. DOI: 10.18653/v1/P16-1030. URL: https://aclanthology.org/P16-1030.

[108] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. "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)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1650–1659. URL: https://aclanthology.org/P13-1162.

[109] Scott A. Reid. "A Self-Categorization Explanation for the Hostile Media Effect." In: *Journal of Communication* 62.3 (Apr. 2012), pp. 381–399. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.2012.01647.x. eprint: https://academic.oup.com/joc/article-pdf/62/3/381/22321551/jjnlcom0381.pdf. URL: https://doi.org/10.1111/j.1460-2466.2012.01647.x.

[110] Filipe Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna Gummadi. "Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale." In: *Proceedings of the International AAAI Conference on Web and Social Media* 12.1 (July 2018). URL: https://ojs.aaai.org/index.php/ICWSM/article/view/15025.

[111] Dr. John P. Robinson. "Perceived Media Bias and the 1968 Vote: Can the Media Affect Behavior after All?" In: *Journalism Quarterly* 49.2 (1972), pp. 239–246. DOI: 10.1177/107769907204900203. eprint: https://doi.org/10.1177/107769907204900203. URL: https://doi.org/10.1177/107769907204900203.

[112] Axel Rodríguez, Carlos Argueta, and Yi-Ling Chen. "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis." In: *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2019, pp. 169–174. DOI: 10.1109/ICAIIC.2019.8669073.

[113] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. "Social Media News Communities: Gatekeeping, Coverage, and Statement Bias." In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 1679–1684. ISBN: 9781450322638. DOI: 10.1145/2505515.2505623. URL: https://doi.org/10.1145/2505515.2505623.

[114] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. "Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold." In: *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*. 2013. URL: http://oro.open.ac.uk/40660/.

[115] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. "The Risk of Racial Bias in Hate Speech Detection." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1668–1678. DOI: 10.18653/v1/P19-1163. URL: https://aclanthology.org/P19-1163.

[116] Kathleen M. Schmitt, Albert C. Gunther, and Janice L. Liebhart. "Why Partisans See Mass Media as Biased." In: *Communication Research* 31.6 (2004), pp. 623–641. DOI: 10.1177/0093650204269390. eprint: https://doi.org/10.1177/0093650204269390. URL: https://doi.org/10.1177/0093650204269390.

[117] SemEval. *International Workshop on Semantic Evaluation*. URL: https://semeval.github.io/ (visited on 03/07/2022).

[118] Gün R. Semin and Klaus Fiedler. "The cognitive functions of linguistic categories in describing persons: Social cognition and language." In: *Journal of Personality and Social Psychology* 54.4 (1988), pp. 558–568. DOI: https://doi.org/10.1037/0022-3514.54.4.558.

[119] Tom De Smedt, Guy De Pauw, and Pieter Van Ostaeyen. *Automatic Detection of Online Jihadist Hate Speech*. 2018. arXiv: 1803.04596 [cs.CL].

[120] Keith Somerville. "Violence, hate speech and inflammatory broadcasting in Kenya: The problems of definition and identification." In: *Ecquid Novi: African Journalism Studies* 32.1 (2011), pp. 82–101. DOI: 10.1080/02560054.2011.545568. eprint: https://doi.org/10.1080/02560054.2011.545568. URL: https://doi.org/10.1080/02560054.2011.545568.

[121] Daniel Sousa, Luís Sarmento, and Eduarda Mendes Rodrigues. "Characterization of the Twitter @replies Network: Are User Ties Social or Topical?" In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. SMUC '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 63–70. ISBN: 9781450303866. DOI: 10.1145/1871985.1871996. URL: https://doi.org/10.1145/1871985.1871996.

[122] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. "Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias." In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 389–392. ISBN: 9781450375856. DOI: 10.1145/3383583.3398619. URL: https://doi.org/10.1145/3383583.3398619. published.

[123] Timo Spinde, David Krieger, Manu Plank, and Bela Gipp. "Towards A Reliable Ground-Truth For Biased Language Detection." In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. Virtual Event, 2021. DOI: 10.1109/JCDL52503.2021.00053. URL: https://media-bias-research.org/wp-content/uploads/2022/01/Spinde2021d.pdf (visited on 09/01/2021). published.

[124] Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. "Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles." In: *Proceedings of the iConference 2022*. Virtual event, 2022. DOI: https://doi.org/10.1007/978-3-030-96957-8_20. URL: https://media-bias-research.org/wp-content/uploads/2022/03/Spinde2022a_mbg.pdf (visited on 03/04/2022). published.

[125] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. "Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Dominican Republic, 2021. DOI: 10.18653/v1/2021.findings-emnlp.101. URL: https://media-bias-research.org/wp-content/uploads/2022/01/Neural_Media_Bias_Detection_Using_Distant_Supervision_With_BABE___Bias_Annotations_By_Experts_MBG.pdf (visited on 11/01/2021). published.

[126] Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela, Gipp, and Karsten Donnay. "MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics." In: *Proceedings of the iConference 2021*. Beijing, China (Virtual Event), 2021. DOI: 10.6084/m9.figshare.17192924. URL: https://media-bias-research.org/wp-content/uploads/2021/01/MBIC-âĂŞ-A-Media-Bias-Annotation-Dataset-Including-Annotator-Characteristics.pdf (visited on 03/01/2021). published.

[127] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. "Automated identification of bias inducing words in news articles using linguistic and context-oriented features." In: *Information Processing & Management* 58.3 (2021), p. 102505. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102505. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000157/pdfft?md5=64e81212b3bfa861d01a6fe3d5b979c3\&pid=1-s2.0-S0306457321000157-main.pdf. published.

[128] Cass R. Sunstein. "The Law of Group Polarization." In: *John M. Olin Law & Economics Working Paper* 91 (Dec. 1999). DOI: http://dx.doi.org/10.2139/ssrn.199668. URL: https://dash.harvard.edu/handle/1/13030952.

[129] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. "Sentiment strength detection for the social web." In: *Journal of the American Society for Information Science and Technology* 63.1 (2012), pp. 163–173. DOI: https://doi.org/10.1002/asi.21662. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21662. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21662.

[130] Joel Turner. "The Messenger Overwhelming the Message: Ideological Cues and Perceptions of Bias in Television News." In: *Political Behavior* 29 (2007), pp. 441–464. DOI: 10.1007/s11109-007-9031-z. URL: https://doi.org/10.1007/s11109-007-9031-z.

[131] Robert P. Vallone, Lee Ross, and Mark R. Lepper. "The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre." In: *Journal of Personality and Social Psychology* 49.3 (1985), pp. 557–585. DOI: https://doi.org/10.1037/0022-3514.49.3.577.

[132] Claes H. de Vreese. "News framing: Theory and typology." In: *Information Design Journal* 13.1 (2005), pp. 51–62. ISSN: 0142-5471. DOI: https://doi.org/10.1075/idjdd.13.1.06vre. URL: https://www.jbe-platform.com/content/journals/10.1075/idjdd.13.1.06vre.

[133] Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 225–230. DOI: 10.18653/v1/P16-2037. URL: https://aclanthology.org/P16-2037.

[134] William Warner and Julia Hirschberg. "Detecting Hate Speech on the World Wide Web." In: *Proceedings of the Second Workshop on Language in Social Media*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 19–26. URL: https://aclanthology.org/W12-2103.

[135] Zeerak Waseem. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 138–142. DOI: 10.18653/v1/W16-5618. URL: https://aclanthology.org/W16-5618.

[136] Zeerak Waseem and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. DOI: 10.18653/v1/N16-2013. URL: https://aclanthology.org/N16-2013.

[137] Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. "Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4090–4105. DOI: 10.18653/v1/2020.emnlp-main.335. URL: https://aclanthology.org/2020.emnlp-main.335.

[138] David Manning White. "The "Gate Keeper": A Case Study in the Selection of News." In: *Journalism Quarterly* 27.4 (1950), pp. 383–390. DOI: 10.1177/107769905002700403. eprint: https://doi.org/10.1177/107769905002700403. URL: https://doi.org/10.1177/107769905002700403.

[139] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. URL: https://aclanthology.org/H05-1044.

[140] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[141] Xue Ying. "An Overview of Overfitting and its Solutions." In: *Journal of Physics: Conference Series* 1168.2 (Feb. 2019), pp. 1–7. DOI: 10.1088/1742-6596/1168/2/022022. URL: https://doi.org/10.1088/1742-6596/1168/2/022022.

[142] Gi Woong Yun, Sung-Yeon Park, Sooyoung Lee, and Mark A. Flynn. "Hostile Media or Hostile Source? Bias Perception of Shared News." In: *Social Science Computer Review* 36.1 (2018), pp. 21–35. DOI: 10.1177/0894439316684481. eprint: https://doi.org/10.1177/0894439316684481. URL: https://doi.org/10.1177/0894439316684481.

[143] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. "Measuring and Characterizing Hate Speech on News Websites." In: *12th ACM Conference on Web Science*. WebSci '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 125–134. ISBN: 9781450379892. DOI: 10.1145/3394231.3397902. URL: https://doi.org/10.1145/3394231.3397902.

[144] Xueying Zhang and Mei-Chen Lin. "The Effects of Social Identities and Issue Involvement on Perceptions of Media Bias Against Gun Owners and Intention to Participate in Discursive Activities: In the Context of the Media Coverage of Mass Shootings." In: *Mass Communication and Society* 25.2 (2022), pp. 260–281. DOI: 10.1080/15205436.2021.1916036. eprint: https://doi.org/10.1080/15205436.2021.1916036. URL: https://doi.org/10.1080/15205436.2021.1916036.

[145] Ziqi Zhang and Le Luo. "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter." In: *Semantic Web* 10.5 (2019), pp. 925–945.

[146] Ziqi Zhang, David Robinson, and Jonathan Tepper. "Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network." In: *Proceedings of ACM The Web conference (WWW'2018)* 4 (2018).

[147] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation." In: *ACM Transactions on Management Information Systems (TMIS)* 9.2 (Aug. 2018). ISSN: 2158-656X. DOI: 10.1145/3185045. URL: https://doi.org/10.1145/3185045.