COPENHAGEN BUSINESS SCHOOL

*A thesis presented in fulfillment of the requirements for a*

**MSc. in Business Administration and Data Science**

# It's 2023 and Women are Still "Hysterical": Analyzing Bias in BBC News

## Has Gender Bias in Language Used in the Media Changed In the Age of Datafication?

Danielle Verniece Duncan (150419)

Yolanda Valentina Ferreiro Franchi (149102)

Supervisor: Dr. Daniel Hardt

Pages: 108

Characters: 166,007

Submitted May 14, 2023

# Abstract

The rise of datafication has transformed the news industry by modernizing it from a daily paper release to real-time updates on the internet. This expansion of the news has enabled it to reach consumers more easily than ever before. The news as an institution is seen as a source of unbiased truth, however this is not necessarily the case. As the news expands its reach and coverage, ensuring that reporting is fair and unbiased is more important than ever. This research presents an approach to detecting gender bias in news text by utilizing logistic regression and Bi-LSTM RNN machine learning models.

Gender bias detection is a difficult task because it is intangible and hard to codify. As society grows and changes, the bias evolves with it. Our study reveals that while the representation of women in news reporting has increased over the years, bias is still present. This bias is evident through the ability to accurately predict if a sentence is male or female with gender identifiers removed. Our analysis of news data reveals that many words stereotypically associated with genders are strongly predictive of those genders. For example, ”bossy” is highly predictive of the female class, and ”leader” for the masculine class.

Our findings suggest that the BBC is making efforts to become less biased, but more adjustments need to be made. We also highlight the significance of our work, emphasizing how implicit bias can feed explicit bias, which can then perpetuate stereotypes in society. Our study contributes to the growing field of natural language processing for social good, demonstrating the potential of machine learning techniques to uncover and address bias in written text.

**Keywords:** Bi-LSTM RNN; Gender Bias; Logistic Regression; Media Text; Natural Language Processing; NLP for Social Good; Text Classification

# A Note from the Authors:

All the code in this paper can be found on Github:

https://github.com/rainbowjoy1/Masters-Thesis.

The code is provided in Jupyter Notebooks utilizing Python. Graphs were made using PowerBI and the dataset can be provided for validation purposes upon request. If you would like to explore the data in a dynamic way, and interact with the coefficients and weights, refer to this study's PowerBI Dashboard: https://bit.ly/3NGLwVl

This thesis examines complex and intricate systems of gender inequality. While care has been taken to include the most up-to-date theories, important aspects have been left out. This study views gender on the male/female binary. The authors acknowledge that this neglects individuals identifying as non-binary. Though this is a common approach in machine learning, the authors feel it is important to highlight the lack of representation of certain individuals due to this choice, as this paper spotlights a need for more representation. While this is an understandable choice, it is still an glaring omission.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

Part I

# INTRODUCTION

# 1

# BACKGROUND

Gender is a phenomenon of undeniable cultural impact. Sociologists say that it is one of the earliest intergroup differences that humans are able to observe (Goodhew et al. 2022). Despite the progress made in the past century, gender bias remains a pervasive social issue globally, regardless of economic and political development. While a tale as old as time, it has only recently been given a name due to the first and second waves of feminism. Gender bias continues to exist in various aspects of society, including The Media. Natural language is a powerful tool that can reflect and perpetuate gender bias. Analyzing language therefore is a key step towards addressing this issue.

Natural language processing (NLP) has gained attention in recent years for its ability to examine and comprehend human language. An area where its techniques have been successfully applied is text classification where they have yielded highly accurate results with strong plausibility. This paper frames the issue of identifying gender bias in the news as a text classification problem. Employing an intersectional approach, our research combines data and social science to build text analysis models. We analyze results through the lens of feminist linguistics, to detect bias based on how individuals who suffer it see it and describe it. The impetus for this undertaking is an aspiration to reduce gendered prejudice from the fabric of human society, as bias of any kind is fundamentally incompatible with fairness and inclusivity. This line of research aligns with the emerging ”NLP for Social Good” movement in machine learning, which seeks to inspire researchers to amplify the ethical and societal impact of their work.

With the rise of datafication, the reach of the news institutions has proliferated. The Media is thus an ideal institution in which to study the evolution of gender bias over time as it generates large amounts of natural language data. Using machine learning algorithms, said data can be analyzed to identify patterns that are not humanly apparent. This study focuses on BBC News over a period of 12 years, employing both a traditional and deep learning NLP approach to investigate how gender bias manifests. The paper seeks to capture and quantify explicit and implicit forms of bias during this time. The hope is that by identifying how gender bias surfaces in the news, the necessary adjustments can be made to eschew it in the future.

# 2

# MOTIVATION

Gender bias is a persistent social issue because it is perpetuated by the institutions that make up the social system. The Media is one of these institutions. It is of pervasive influence in shaping values, beliefs, and attitudes due to the function it fulfills: to inform. Its reach and impact have been amplified in the current age of datafication, where there has been a proliferation of data in all parts of human life.

Media institutions have a moral responsibility to ensure that news reporting is truthful, accurate and impartial. News reporting exhibiting gender bias - or any bias for that matter - undermines this objective. By analyzing news text for gender bias, media institutions can begin to understand how they perpetuate said bias and what they can change to make their reporting more impartial. This process, in turn, can help foster trust with their audience, as readers lean towards engaging with news institutions that they perceive as objective and fair.

Datafication has prompted the adoption of machine learning. Several sub-fields of the latter such as NLP have seen a meteoric rise in application. NLP combines machine learning algorithms with computational linguistics to automate the process of analyzing written natural language text. Among other things, it provides a potent collection of methodologies that can be used to identify bias in text. NLP algorithms - regardless of whether statistical, machine learning or deep learning focused - enable one to efficiently process large amounts of text data, simulating human ability and understanding. These methods are capable of capturing difficult or near impossible patterns to

detect manually. In the context of this application, these capabilities are useful given that bias is not always manifested explicitly, and often lies in implicit language choices.

Gender bias in the news can reinforce harmful attitudes that contribute to the marginalization of groups that are not a part of the dominant social class. By leveraging NLP methodologies to identify bias in news text, a more nuanced, scientific understanding of how gender is represented can be gained. Once bias is identified, the necessary steps to ensure better reporting can concretely be taken. The power of this approach when it comes to mitigating gender bias is that it would be driven by tangible, algorithmic insights. Thus, it would tackle the issue objectively rather subjectively as most current inclusivity initiatives do.

# 3

# PROBLEM FORMULATION

Women face large and varied amounts of bias based on the notions of traditional gender roles. This bias manifests itself through social institutions in many mediums, including written text. The Media has been an institution of great influence over time, responsible for keeping people informed. However, it has not always lived up to its moral duty of reporting in a fair and impartial manner. In fact, throughout time, it has represented and portrayed individuals stereotypically. From a business viewpoint, this is understandable given that it seeks to connect with audiences through assumed shared understandings of identity (Goodhew et al. 2022). For instance, women are frequently portrayed in a feminine light: as emotional, vain, and overall more concerned with domestic life. Men, on the other hand, tend to be characterized as assertive, ambitious, and more worried about work life. These characterizations have a strong social impact, such as perpetuating biased notions of gender. These consequences are significant because they are completely baseless in reality, given that women possess intellectual and interpersonal attributes which make them equally as capable - if not more so in some scenarios - than men.

NLP has gained recognition in recent years for its ability to generate insights from written text, identifying patterns in language that are not obvious to humans. It has a proven track-record in business applications, specifically in the areas of sentiment analysis, text classification, and named entity recognition. Yet, socially-focused issues such as bias are under-explored in NLP scholarship. Moreover, the limited literature in NLP on gender bias tends to be focused on mitigating it, more so than identifying it. The latter is precisely the area in which this study seeks to contribute to the

existing scholarship. The pretense for this is that to effectively mitigate gender bias, one first needs to holistically comprehend how it is constituted in text-effective mitigation is not possible without proper identification.

Identifying gender bias in language is no easy feat. It is challenging for individuals because it can be heavily codified and sometimes subjective. Recognizing gender bias automatically and accurately through NLP models is also an arduous task as there are various ways to capture linguistic patterns that evince it but they must be codified. Given this, an intersectional approach which evaluates the results of NLP models from the lens of feminist linguistics is necessary to capture the full spectrum of gender bias. Similar approaches are currently lacking in NLP literature but are necessary to capture all of the nuances and dimensions of bias, particularly as seen by individuals against whom it is targeted towards.

Systematic change is occurring globally, particularly pertaining to issues of equality and inclusivity. This transformation has taken place due to evolving attitudes on social issues, including gender equality. Datafication has played a big role in sparking change by increasing the availability and volume of information, enabling individuals to forge generally more progressive attitudes. Evolving attitudes are apparent in legislation and social movements, but less obvious when it comes to language. This brings forth the question this thesis seeks to answer.

## 3.1 THESIS SCOPE

Combining all of the points in the problem formulation, this thesis will investigate whether language used in media reporting has evolved over time when it comes to gender bias.

This thesis utilizes the British Broadcasting Corporation (BBC) news as a proxy for media overall because it is one of the top five news institutions in the world by audience. For perspective, in 2021, it had a weekly reach of 186 million adults and an audience of over 500 million people per year (News 2021). While conclusions about the BBC cannot speak for all news agencies, the breadth of their audience, size of their newsroom, and number of news articles published makes them a pivotal player in the role of The Media in people's lives.

An approach leveraging traditional and deep learning NLP techniques will be utilized to analyze different aspects of the evolution of language. These methods seek to identify patterns of gender bias at the word and sentence levels of BBC news text over a period of 12 years: from 2010 to 2022. The scope of gender bias analyzed will be limited to the following categories discussed in feminist linguistics: power, violence, appearance, communality, agency, and representation.

It is worth noting that the BBC has signalled interest in issues of gender inequality. In 2017, it launched the award-winning 50:50 project with the intention of having equal representation in their editorial staff engaged in article production. This 50:50 rule does not apply to quoted individuals as they ”cannot tell the stories without these people, and we have no control over who they are” (News 2023). While a step forwards, the 50:50 project is an example of a current subjective initiative used to tackle inclusivity. For the BBC to drive the institutional change required to thrive in an increasingly progressive social climate, a more objective initiative is required. This is because equal representation in terms of authorship might not necessarily be related to the quality of representation of individuals. The research herein hopes to highlight how it can improve in terms of the latter.

## 3.2 RESEARCH QUESTION

This thesis aims to answer the research question (RQ) based on the aforementioned scope. To comprehensively answer said RQ, the hypotheses in the table below will be considered:

| | |
|---|---|
| **RQ** | Has Gender Bias in Language Used in the Media Changed In the Age of Datafication? |
| **H1** | Diction choices differs in sentences where the subject is male or female. |
| **H2** | Syntactic choices change in sentences where the subject of the sentence is male or female. |
| **H3** | Bias is exhibited and explainable in sentences where the subject is male or female. |

Table 1: Research Question and Hypotheses

## 3.3 DEFINITIONS RELATED TO THE RESEARCH QUESTION

To ensure that the components of this study are understood in the way the authors intend them to, the following definitions will be utilized.

| Term | Definition | Source |
|---|---|---|
| **Diction** | Concept referring to the choice of words to express meaning in a text. | Halliday 1985 |
| **Syntax** | The rules that dictate the order of words and the use of grammatical structures. | Chomsky 2002 |
| **Bias** | Systematic favoring of individuals, or groups thereof, in a way that is considered unfair. | Lorber, Farrell, et al. 1991 |

Table 2: Definitions Related to the Research Question

Part II

# CONCEPTUAL FRAMEWORK

# 4

# GENDER

Formal definitions of gender in sociology emerged in the 1930s. Gender was defined in the same vein as MacIver's delineation: "the cultural elaboration of sex differences" (MacIver 1931). Since then, the social and cultural aspects of gender are accepted notions in social science.

An important theme in gender studies are the waves of feminism. These waves historically and socially contextualize the discussions pushed forth by the feminist movement and gender studies scholarship throughout time. To date, there have been four waves of feminism. The first started in the late 19th century and was chiefly concerned with securing political and legal rights for women, such as suffrage and access to education (Fricker and Hornsby 2000). The second emerged in the 1960s, and focused on a broader scope of issues challenging traditional gender roles and cultural norms (Fricker and Hornsby 2000). The third wave, starting in the 1990s, centered around the intersectionality of feminism and different forms of oppression faced by minorities (Schrupp 2017). Lastly, the fourth currently-evolving wave, concerns the empowerment of women and technological activism (Schrupp 2017).

A key milestone of the second wave of feminism was the revision of the definition of gender. During this wave, the fundamental feminist principle that gender is a social construct began to gain traction in social science. This ripple effect began with Ann Oakley's 1972 article "Sex, Gender and Society", which was the first sociological definition of gender incorporating social constructionism. Said notion, however, dates back to Simone de Beauvoir's *The Second Sex* Second Sex, where she proposed that "one is not born, but rather becomes, a woman" (De Beauvoir and Moinaux 1953, p.

267). Since the 1970s, definitions of the term ”gender” have significantly converged across social science as can be observed in the work of various theorists in the multiple fields within it: Rubin 1975, Butler and Trouble 1990, Lorber 1994, and Kimmel et al. 2008.

The significance of the social constructionism element in the definition of gender is that it outlines the mechanics through which the patriarchy has justified the subordination of women. Historically, gender has been contemplated on the male/female binary. This is based on the essentialist view that it is determined by one's sexual organs, and therefore, there are immutable differences between males and females. This discrete categorization is sustained by a "tacit collective agreement" to ”perform” gender (Butler and Trouble 1990, p. 528). More so than being inherent and biological, gender is an act according to social constructionism. It was fabricated by the social system to determine individuals in the dominant and non-dominant class. Gender plays a tremendous role in society and culture - stereotypical performance of it leads to social acceptance and reassurance. Non-conformity leads to the ousting of individuals from their conventional, systemic social groups (Butler and Trouble 1990). Unfortunate realizations of this include: the witch hunts in the 16th and 17th centuries, mistreatment of the suffragettes in the 19th and 20th centuries, and most recently, bias targeted at certain social groups.

Historically, "sex" and "gender" have been used interchangeably. The distinction stems from 1950s psychology, in John Money's works on the psychological management of sex and gender. Money et al. 1960 defines "sex" as being based on five characteristics: "chromosomal sex, gonadal sex, hormonal sex, internal reproductive organs, and external genitalia" and gender as being determined by psychological characteristics and behaviors. Subsequent work in the field further argues that more precision is necessary in demarcating the immutable and mutable elements of both sex and gender, recognizing that psychologically-speaking the former is natural whereas the latter is culturally assumed Unger and Crawford 1993. Thus, the widespread shift in terminology of the second wave of feminism, particularly pertaining on gender, has its roots in psychological theory. This is as opposed to the baseless origin of the term in the patriarchy.

## 4.1 GENDER BIAS

Gender bias is a pervasive social issue globally, regardless of economic and political development. Social sciences study gender bias in relation to its sociocultural implications. Most definitions of the term agree that it refers to the unfair treatment of individuals, or groups thereof, based on their gender (De Beauvoir 1949), (Butler 1988), (Lorber, Farrell, et al. 1991), (Zamudio and Rios 2006). The term is usually applied to denote the systemic unfair treatment of what de Beauvoir's describes as the second sex - women. In recent times, it has begun to represent said against all individuals that are non-stereotype conforming, such trans and non-binary people.

Gender bias exists in many forms. Common ones include stereotyping, prejudice, and discrimination. Gender bias generally has characteristics associated to each respective genders as its relative baseline. Society and culture commonly stereotypes personality attributes, occupations, activities, and romantic roles (Ward and Grower 2020). This results in: behaviorally, females being portrayed as "communal"; occupationally, as housewives or in occupations aligning with their gender stereotypes, such as nurse or teacher; and romantically, as overly emotional (Ward and Grower 2020). Male counterparts are generally associated with being more verbally and physically aggressive, occupying themselves in the world of work, and romantically, as ignoring their emotions while dictating courtship.

Within the context of the social system, an important notion brought forth by feminists is that gender bias is a self-reinforcing cycle. In the 1980s, bell hooks was the first to highlight that gender bias is what leads to the socioeconomic inequality between men and women. She argues that this practice is embedded in social structures including cultural norms, politics, and public institutions. Lorber further holds that gender bias is sustained by the constantly evolving social practice of gender (Lorber 2001). This negative feedback loop holds in public life (Crenshaw 1989), business organizational practices (Acker 1992); and the legal system (Fraser 1990).

# 5

# SYSTEMS OF OPPRESSION

Modern society occupies a complex space balancing individual behaviors with the context of institutions and systems. Systems are "any collection of interrelated parts or elements that we can think of as whole" (Johnson 2004). These systems, while they cannot exist without us, do not exist because of us. Johnson brings up the example of the corporation. Everyone that works for the corporation is part of the corporation but if everyone were to quit the corporation, they could be replaced and the corporation would continue to run. This complexity of systems, being a part of people but also external to them, makes them complex to work with and change. These systems are seen on minor scales like companies but on macro scales ranging from governmental systems to capitalism as a whole concept. Systems, while not living or breathing, are beings that have a life of their own that are propped up through the behaviors and values that the system deems "appropriate and expected" with rules that are "external and beyond our control" (Johnson 2004).

These systems also include social systems that generate systems of oppression. Systems of oppression are systemic, directional power relationships among social identity groups, in which one group benefits at the expense of other groups (Adams and Bell 2007). Social oppression is important to define because it exists outside of purely individualistic behaviors and views and is more than one group's assertion of superiority. Adams and Bell (2007) define social oppression as existing when the following conditions are met: the agent group defines what is "normal", "real" or "correct"; differential treatment is institutionalized and systematic; the behavior is systematic and rote; the oppressed internalize their oppression and collude with their oppressors; the target's

culture is removed or repressed and the dominant culture is imposed. These paradigms of social oppression are part of racism, ableism, heterosexualism, sexism, and more.

The overarching social system of the patriarchy is seen as an encapsulation of the social system of systematic sexism. The system of patriarchy is woven into the fabric of society through social, governmental, and cultural systems. The identification of the patriarchy is often attributed to Engels who described it as concentration of power towards men through the control of property and goods. Feminist historians argued that "state, class, and patriarchy arose together" attributing that the entire patriarchal structure is due to the economic exploitation of class and the rise of the state, making clear that the patriarchy has been an aspect of society for most of discernible history (Omvedt 1987). This identification of the historical presence of patriarchy led feminists to continue the development of the definition and show that this social system continues to influence behavior to present times.

Most common definitions see the patriarchy as a system that is male-dominated, male-identified, and male-centered (Johnson 2004), (Christ 2016), (Facio 2018). Because the patriarchy is a system, its influence is far-reaching by encompassing language, policy, appearance, and entire conceptualizations of self. The patriarchy is about how life is and how it is supposed to be. For further defining characteristics of the patriarchy see Section 5.2. By centering men and male-ness as the defining characteristic of the system, anything that is outside the scope of male is aberrant and must be quelled and controlled. This narrative of aberrance and control leads the patriarchy into using various methods of oppression because the system must self-perpetuate.

Young (1990) introduced the five faces of oppression to highlight that oppression has different faces depending on the agent and victim in the relationship. Some methods of oppression are more codified like exploitation and marginalization with laws and policies often supporting the oppression. Others like powerlessness and cultural imperialism are impacted heavily through everyday actions like language choices, group interactions, and internal structures. Some examples would include: calling girls "bossy" but boys "leaders", the finance department being all male and the marketing one all female in a company, claiming dominance with triviality through violent language "I raped that meeting", etc.

| | **Exploitation** | **Cultural Imperialism** | **Powerlessness** | **Marginalization** | **Violence** |
|---|---|---|---|---|---|
| **Definition** | the act of using people's labor to produce profit while not compensating them fairly | taking the culture of the ruling class and establishing it as the norm | inhibition to develop one's capacities, lack of decision making power, and exposure to disrespectful treatment because of the lowered status. | relegating or confining a group of people to a lower social standing or outer limit or edge of society | Members live with the knowledge that they must fear random, unprovoked attacks on their persons or property. Attacks are intended to damage, humiliate, or destroy the person |
| **Institutional Oppressors** | Business groups Government | The Media Government Business groups | Government The Media | Government The Media | Government groups Business |
| **Methods of Oppression** | Withholding pay, illegal labor practices, exploiting legal status/ignorance for financial gain, pay gaps | Gentrification, | Codified | Exclusionary laws, discrimination, | Lynching, rape, hate crimes, hate speech |

Figure 1: 5 Forms of Oppression by Heldke and O'Connor (2004)

These faces of oppression are expressed in different ways depending on the institution engaging in the oppression. While The Media can engage in violent oppression more directly through public calls to action of violence towards outgroups, this is rare compared to the language that the news source uses on a daily basis. While oppression can seem intentional, it is often systematic and unintentional for the individual. This unintentional oppression can be seen most easily through decisions and language choices that groups and individuals make.

## 5.1 LANGUAGE

Language is a powerful tool that humans have been using for thousands of years to communicate with one another. The concept of language is interdisciplinary, playing an integral part of all social science but a different role in each. Sociology considers language as "an interactional tool that we use to shape the world around us and to make sense of it" (Goffman 1981, p. 6). It is seen as a tool to communicate, that is shared by a culture and community, through which they can express themselves and manifest their social behaviors. Linguistics, on the other hand, sees language "a system of generative grammar, a set of rules that allows speakers to produce an infinite number of sentences, each with a unique meaning" (Chomsky 2014, p.3). In this social science, language

is studied more scientifically, and the focus is on the properties of languages more so than its sociocultural impact.

While a hallmark of the human species, the provenance of language has been a topic of debate by researchers. Estimates of when humans started talking to each other range from as early as 2 million to as late as 50,000 years ago (Balter 2015). There are multiple theories that attempt to explain the evolution of spoken language. Müller's "ding-dong" theory proposing that language evolved from a combination of natural sound imitations and hand gestures has been the dominant one over time (Hauser et al. 2014). Following the process of Darwinian natural selection, humans who were equipped with the ability to communicate through language could reap more tangible benefits within a context of limited resources. According to researchers, this process enabled them to gradually improve their primitive language abilities, eventually leading to the "full-blown, semantically complex languages we speak today" (Balter 2015).

A solid body of empirical evidence has shown that language shapes thinking. This idea was developed by linguists in the early 20th century, in what became known as the Sapir-Whorf hypothesis. This hypothesis argues that language can shape our judgments and behaviors due to its relative, or as otherwise known, deterministic quality. In their respective works, Sapir 1929 and Whorf 1956 explain that the language one uses can influence one's and others' thought process and perceptions of the world. It is the key to not only knowledge but also the construction of reality (Boroditsky 2011).

The notion of time is important element when discussing language. The scholarship is resolute that language is a phenomenon which has developed over time, "... through repeated cycles of learning and use" (Chater and Christiansen 2010). At its core, language and like many of the institutions in human society, has historically centered around men. By virtue of its evolution, which occurs by gradually aggregating on itself, language has the potential to reinforce fragments of human reality that are outdated and biased. This can be observed semantically. So while language in itself might be a neutral phenomenon, it can perpetuate stereotypes, prejudices and discrimination depending on its use.

## 5.2 PATRIARCHY

In practice, the patriarchy has key aspects of behavior and linguistic choices that help to support the definition of male-dominated, male-identified, and male-centered. By defining the male as the center, the definition of systems of oppression shows that the female, is therefore not normal, false, and incorrect. The patriarchal system highlights the differences between men and women by valuing masculinity and maleness through devaluing femininity and femaleness (Johnson 2004). For instance, "assertiveness and performance are seen as indicators of greater agency in men, and warmth and care for others are viewed as signs of greater communality in women" (Kite et al. 2008). This agency and communality duality is often coupled with the dichotomy of power and subordination of men and women, where men are seen as harnessing power through governmental roles, family structure, and violence while women are valued for their subordination through religion, the historical lens of them as property, and various other methods. This centers women as a group with lower power, representation, agency, and high communality while the male group has high power, representation, agency, and low communality.

# 6

# MACHINE LEARNING

## 6.1 NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is the branch of machine learning concerned with the analysis of natural language text. It uses computer to process language like a human, developing algorithms that can be employed to understand, interpret, and generate natural language data such as text, speech, and gestures.

A computer is a complex arrangement of switches that take inputs of ones and zeros (on and off) to represent more complex systems. This is seen in simple applications like math where numbers are encoded with binary (a series of switches) and transistors are applied to enable processes like addition. When language is used, each letter has a number that corresponds to it that is encoded in unicode and then to binary. Each letter, upper and lowercase, has its own number that corresponds to it. This is an incredibly naive understanding of language as the computer's internal representation of "Hello", "hello", "hi", and a waving emoji are incredibly different internally despite that from a linguistic perspective these words are close to identical.

NLP emerged in the 1950s, when researchers began exploring how to apply computers to process and understand the human language. The intent was to create machines that could pass the Turing Test - ergo, have the ability to communicate humanly. Weaver's 1955 "A Statistical Approach to Machine Translation" was one of the earliest works in NLP. In this paper, he proposed that language, at its heart, is a statistical phenomenon around which algorithms could be built to analyze large

amounts of data with the ultimate intent of comprehending and creating natural language (Weaver 1955). Weaver's approach is the foundation of many of the statistical techniques and algorithms nowadays. Important applications of his foundation are featured in part-of-speech tagging, named entity recognition, machine translation, and sentiment analysis.

The age of datafication has led to an increased adoption of machine learning in both business and social institutions. This is what has led to an explosion in NLP applications. NLP techniques have had a positive track record in: text summarization and classification: (P. Liang 2005), (Kalchbrenner et al. 2014), (Goldberg 2017); sentiment analysis: (Go et al. 2009), (Pankajakshan et al. 2021); named entity recognition: (Chiu and Nichols 2016); topic modeling:(Blei et al. 2003), (Alvarez-Melis and Saveski 2016); speech recognition (Graves and Schmidhuber 2005), and information retrieval (Mikolov et al. 2013), (H. Zhang et al. 2021), among others. These techniques have transformed organizational operations, leading to higher efficiency and profitability. Mainstream examples of NLP applications that have transformed value propositions and competition include: information retrieval for customer service chatbots and search engines (Rahimi 2019); sentiment analysis for market research, capturing customer feedback, and reputation management (Høysæter and Njølstad 2014), (Zuheros et al. 2021); using text classification for spam filtering (Metsis et al. 2006); and named entity recognition for financial analysis (Murtagh 2018).

### 6.1.1 *Term Frequency-Inverse Document Frequency*

Term frequency (TF) represents the number of times a particular term appears in a document divided by the sum of the words in said document. This concept was introduced by Salton and Buckley in 1988 in their influential work on term frequency and text retrieval. The authors proved that by incorporating term frequency information into text retrieval processes, the performance of information retrieval systems significantly improved. Log space is used to deal with large collections of documents (Jurafsky and Martin 2021). This space is useful as it is used to ensure that a word that appears a thousand times is not considered a thousand times more important than a word appearing just once.

In addition to proposing the raw term frequency scheme, Salton and Buckley also brought forth term frequency-inverse document frequency (TF-IDF). TF-IDF - as per the formula in Appendix 18.5.1 - includes an additional component to term frequency that considers the rarity of a word in the corpus as a whole - that of "inverse document frequency" (IDF). The idea of IDF emerged in the 1970s in Sparck Jones (1972) and Salton (1975) where it is introduced as a weighting scheme for information retrieval in a vector space. This IDF component is useful because words like "to" or "and" are not as informative as other less frequent words. Therefore, using the TF-IDF scheme, their importance can be weighed down using a lower weight for frequent terms across a collection of documents as they are assumed to carry more specific meaning or context.

TF-IDF is a heavily-used technique to encode text in machine learning. It converts a collection of raw text documents into a matrix of numerical features. In this matrix, each row represents a document, and each column constitutes a unique word in the corpus. The encoding process begins with the tokenization of the text. The next steps in the encoding process are nicely captured by the TF-IDF formula in Appendix 18.5.1. The tf_ij element represents the determined frequency of each token in each document, resulting in the term frequency matrix. Said TF matrix is then multiplied by the IDF weights for each token, whereby the IDF weight of a token is calculated using a logarithm of the total number of documents divided by the number of documents containing said token. The output of this process is a matrix of TF-IDF values which can be used as an input to machine learning algorithms to predict target labels based on the encoded text. Having the ability to assign a higher weight to rare and informative terms, and lower weights to common, less informative ones makes TF-IDF a good method for NLP applications such as information retrieval (Sparck Jones 1972), (H. Zhang et al. 2021); keyword extraction (Ramos 2003), (Wei et al. 2021); sentiment analysis (Akhtar et al. 2017), (Pankajakshan et al. 2021); topic modeling (Vangara et al. 2020), and text classification (Yang and X. Liu 1999), among others.

### 6.1.2 *Sentiment Analysis*

Sentiment analysis (SA) - also known as "opinion mining" - is a NLP technique used to extract and examine information from text. The purpose of this technique is to capture people's opinions, sentiments, beliefs, attitudes, and perceptions (Birjali et al. 2021). SA is a potent tool to gather insights that can be utilized for better decision-making. In the age of e-commerce, where there is a fast evolution of internet-based businesses and applications, SA has gained popularity in various contexts and applications. In business, SA is used to analyze customer feedback on products and customer service (B. Liu 2012); monitor brands (Rasool et al. 2019); manage reputation (Chiranjeevi et al. 2019); categorize online reviews (Fang and Zhan 2015); predict stock prices and market trends (Ahnve et al. 2020); and maximize social media initiatives (Drus and Khalid 2019).

Various techniques can be used to assess the sentiment in a text. Common methods are based on: lexicons, machine-learning, and deep-learning. The first relies on sentiment lexicons that contain pre-assigned sentiment scores (B. Liu 2012); the second detects sentiment through algorithms (Bo and Lee 2008), and the third leverages neural networks for text classification (Socher et al. 2013). Supervised machine learning algorithms - such as Naive Bayes and Support Vector Machine - are the most used techniques in this field for their simplicity and accuracy (Birjali et al. 2021). Due to the absence of labeled data in a lot of SA applications, reinforcement learning techniques are on the rise in the field.

Recent literature highlights how NLP has transitioned from a mostly theoretical field to one of real-world application. Among these applications is that of driving positive social impact, in a trend that is coined as "NLP for social good" (Hovy et al. 2017); (Jin 2021). Researchers argue that among the NLP techniques that can be used to this end is SA, due to its potential to quantify emotions related to different social issues (Mabokela and Schlippe 2022). On the flip side of the coin, the literature also suggests that SA is prone to bias. Kiritchenko and Mohammad (2018) show in their study of SA systems, that many SA systems show statistically significant bias, "providing higher sentiment intensity predictions for one race or one gender". In addition to this, from a computational perspective, SA is limited when it comes to the interpretation of context-dependent

language (Tubishat et al. 2018). This entails that there is still a long way to go before all the relevant sentiment in everyday language, such as the one expressed implicitly through irony, sarcasm, and metaphoric language, can be identified.

### 6.1.3 *Logistic Regression*

Logistic Regression (LR) is a supervised, statistical model widely used for classification tasks in machine learning. It is a linear classification algorithm, and it is used to model both binary and multi-class outcomes. LR - whose formula may be found in Appendix 18.5.2 - operates by applying a logistic function to the input data and passing the result through a sigmoidal activation function (Grus 2019). In this function, as the input gets large and positive, it gets closer to the value of 1, and as it gets large and small, it gets closer to 0 (Grus, 2019). The model is ultimately fit to the data by using gradient descent to maximize the likelihood of the data (Grus 2019). The output of LR represents the probability of the input belonging to a particular category.

LR is commonly used for text classification problems in NLP. This algorithm has achieved importance in machine learning due to ability generate good results albeit its relative simplicity (Shah et al. 2020). Important applications of LR include: fraud detection (Itoo and Singh 2021); marketing to predict customer purchases (Yeung and Yee 2011); risk assessment (Yurynets et al. 2019); predicting customer churn (Markapudi et al. 2021) and employee turnover (Setiawan et al. 2020); and quality control (Kazar et al. 2022), among others. On top of its performance, LR is well-suited for many application due to its computational efficiency and its ability to handle a large number of predictor variables swiftly.

Research on NLP techniques for text classification has repeatedly shown LR to be a stronger performer compared to other traditional machine learning classifiers. Shah et al. (2020) highlight how LR attains the highest precision, accuracy, and F1-score in classifying different categories of text compared to other commonly used classifiers such as Random Forest, and K-nearest neighbors. Pranckevičius and Marcinkevičius (2017) have a similar conclusion when comparing LR to Naive Bayes and Decision Trees for review classification. The literature also suggests LR performs as

well, if not better than, support vector machine (SVM) for text-classification problems (Wang et al. 2020), (Wendland et al. 2021).

### 6.1.4 *Deep Learning*

Deep learning (DL) is the collective name given to a family of algorithms known as artificial neural networks (Muller and Guido 2016). It is a trendy and emerging branch within machine learning, developed to address the limitations of conventional machine learning methods, particularly when it comes to abstract, unstructured data. The premise of DL is the ability to devise models that are composed of multiple processing layers that can learn complex relationships, non-linear representations of data (LeCun et al. 2015). DL involves the use of artificial neural networks - often called neural networks (NNs), for short - to gain intelligence. NNs are mathematical frameworks, inspired by the human brain, which consist of layers of interconnected neurons that process and transform data from an input vector space to an output vector space (Dong et al. 2021). Each artificial neuron - also known as nodes - is connected to other nodes via weights, which determine the strength of the connections, and ultimately, the output of each node.

Technological advancements in the late 2000s led to a resurgence of interest in DL, which was originally conceptualized in the 1940s. This was driven by increased amounts of data and computational power available in the age of datafication (Hao et al. 2016). Recently, DL has gained attention in various fields and shows great promise in many machine learning applications. Concrete examples include natural language processing (Kamath et al. 2019), image and video recognition (Jing and Tian 2020), object detection (Zaidi et al. 2022), speech recognition (LeCun et al. 2015), and robotics (Komal Aggarwal et al. 2022). DL models tend to be carefully tailored to the use case they are applied to, since performance is highly dependent on the specific problem (Muller and Guido 2016). They are computationally expensive to run due to their multiple layers and the volume of data they need to be adequately trained and optimized (Dong et al. 2021).

Complexity in DL algorithms is based on the number of layers, their type, and the parameters involved. There are three main types of DL architectures: feedforward neural networks, convo-

lutional neural networks (CNN), and recurrent neural networks (RNN). The first, is the simplest architecture with uniquely information flowing in one direction. The second is more complex than the first, and consists of convolutional, pooling, and fully connected layers (Vasilev et al. 2019). Finally, RNNs are the most complex type of DL architecture as they contain loops which enables them to have a memory component (Vasilev et al. 2019).

DL architectures differ from traditional machine learning in the following ways: they consist of multiple layers, are composed of interconnected neurons, and learn from data through back-propagation. For data small in size, DL algorithms can perform like traditional classifiers such as LR and SVM as its neurons are essentially classifying nodes (Dong et al. 2021). The full power of DL is realized when dealing with large amounts of data, as they are capable of modeling complex relationships that simpler, traditional machine learning methods cannot (Dong et al. 2021).

### 6.1.5 *Recurrent Neural Networks*

Recurrent Neural Networks (RNN) are a type of neural network specialized for the processing of sequential data, such as text and time series (Muller and Guido 2016). The vast body of NLP research employing RNNs for text applications, classification particularly, is evidence of this (Yu et al. 2019), (Moghar and Hamiche 2020), (Kashid et al. 2023). A supervised deep learning approach, it is well-known for its ability to produce highly interpretable outputs. Contrary to most approaches, RNNs can produce text-based outputs instead of just assigning a label or value for an input (Yu et al. 2019).

Unlike conventional neural networks, which process inputs independently, RNNs are architecturally distinguished by their memory component. Through its layers, they are able to develop a short-term memory, by learning from what has been observed in prior and subsequent layers (Kashid et al. 2023). As such, they can process information in a way that considers the temporal component of the data. This makes them well suited for text-related tasks such as language modelling, sentiment analysis, and text classification.

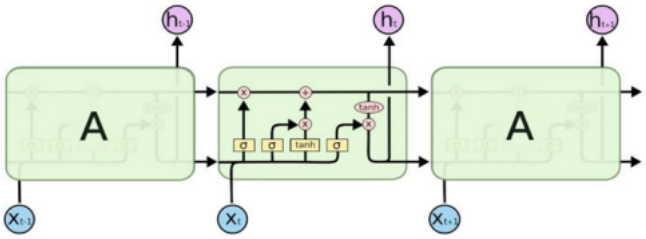


Figure 2: Internal Gate Structure of LSTM by Moghar and Hamiche (2020)

There are multiple types of RNNs. Long short-term memory (LSTM) is an example of such. LSTM is frequently employed in RNN architectures to further extend its memory. Said networks employ memory cells, to selectively store and access information. These cells are regulated by three gates - the forget, memory and output gates (Moghar and Hamiche 2020), which control the flow of intelligence, as represented in Figure 2. The purpose of these gates is to control the state of each gate. Intelligence in LSTM RNNs is thus managed selectively, in a lean manner, and considering the output that should be generated for each time step by virtue of the forget, memory, and output gates respectively. It is through these gates that a key limitation of RNNs is overcome: the inability to store long-time memory (Moghar and Hamiche 2020). Instead, leveraging LSTM networks via the memory cells are able to recognize patterns and relationships that span across multiple time stages, making them more sophisticated than regular RNNs (Moghar and Hamiche 2020).

Part III

# LITERATURE REVIEW

# 7

# LITERATURE REVIEW

## 7.1 LANGUAGE AND FEMINISM

Language and feminism have had a complex history over time. Their relationship is significant because it explains the quality and volume of female representation, in this case, in the news. The Sapir-Whorf hypothesis holds that language shapes how people perceive and think about the world. This is because language is the chief vehicle of expression through which associations among individuals emerge. Consequentially, however, it is also the vehicle through which humans categorize themselves and the stereotypes associated to them based on the observable difference they recognize amongst groups (Goodhew et al. 2022).

While language in and of itself is conventionally-accepted as a neutral instrument, some languages distinguish between genders structurally, for instance, through grammar. Linguistics propounds that three types of languages exist: genderless, natural gender, and gendered languages (Prewitt-Freilino et al. 2012). All of them mark gender using a binary system of categorization. English is natural gender language as most nouns do not have a grammatical marking of gender, like in gendered ones (Prewitt-Freilino et al. 2012). In natural gender languages, gender is indicated using gendered nouns, third person pronouns, occupational and honorific titles, first names, and lastly, the generic "he" (Bigler and Leaper 2015). Socially, these criteria serve the purpose of distinguishing between participants of different groups. From the lenses of feminist linguistics, they delineate individuals in the dominant and non-dominant class within the patriarchy. They also detail the fact that the

generic human existence is masculine. Rhetorical terms like ”he” and ”man”, which are commonly used to describe groups of people, are constitutionally male Bailey et al. (2022). The imagery this relays is that people are men. Therefore, while language might be neutral, the cultural use and implications of it propounds are not.

The second wave of feminism in the 1960s and 1970s brought forth issues around gender-based inequalities in society. One of its most important milestones was accomplished due to feminist linguistics. Feminist linguistics argues that language is not only a communicative tool but also one of repression. This justification was utilized at the time to revisit important terminology, such as the term ”gender” which was redefined to include social constructionism. This idea emerged because it became acknowledged that the use of language was not neutral but rather shaped by social and cultural forces. In a social context, language has the power to reinforce constructed social identities and could thus be considered a tool for oppression as it substantiated the patriarchy (McConnell-Ginet 2014).

An important theme in feminist linguistics is representation. Women are both under- and mis-represented in society, and therefore in language and instances of it, like The Media. They are not represented as much as men, and they are not characterized substantially nor descriptively. Their depictions are stereotype-conforming, and not entirely reflective of the full spectrum of their person. Feminist linguists highlight that descriptors are an indicative element of the quality of female portrayal. Commonly-used descriptors of women include: domesticity (Friedan 1963); an obsession with their appearance and beauty (Wolf 2013); personality attributes such as being weak, emotional, caring (Lawson et al. 2022); romantic traits such as being sexually promiscuous (Ward and Grower 2020); and of relatively lesser capability than men. Feminist scholarship argues that these descriptions seek to articulate a clear distinction between men and women through language. They originate from patriarchal systems of oppression and aim to preserve the status quo. Succeeding at this, necessitates the under-and mis-representation of non-dominant classes. This is because by making them powerless and marginalized, the cultural supremacy of the dominant

class is maintained, thus keeping all aspects of life predominantly androcentric (De Beauvoir and Moinaux 1953).

Under-representation may be quantified in multiple ways. It occurs in terms of volume and quality. Women are often portrayed in ways that reinforce gender stereotypes. In fact, non-conventional perspectives of gender are often dismissed. This results in a relatively smaller share of voices showcasing varied and complex portrayals of women in mainstream culture (Rowe-Finkbeiner 2004). Under-representation is a real issue not only in terms of characters, but also content creators. Women are under-represented in many areas of socioeconomic life including: the news (Asr et al. 2021), cinema (Kagan et al. 2020), television (French 2014), and the legal, political, and science fields (Teele et al. 2018).

## 7.2 DETECTING BIAS

When trying to detect bias it is important to consider both explicit and implicit forms of it. Implicit bias is more codified than explicit bias. The latter can be the presence of racial slurs, abusive language, and exclusionary language. It is more easily identifiable by simplistic methods like word counts and cosine similarity. Per contra, implicit bias often manifest in subtle and nuanced ways Formanowicz and Hansen 2022. Recasens et al. 2013 highlight that automated bias detectors are hard to develop because tracking down sources of bias is hard even for humans. Specifically, for implicit bias, there is an "epistemological" quality to it that makes it hard to identify (Recasens et al., 2016). This quality is exhibited when language choice can subtly alter the believability of a proposition. Recasens et al. (2013) give the example of "claimed" versus "stated", claiming the latter removes the bias introduced by the first word choice, which inherently casts doubt on a statement as it is less assertive. They claim algorithms are not yet well-suited to pick up on these nuances which make all the difference to humans linguistically.

NLP can be effective at identifying certain types of biases in text. It is particularly good at detecting explicit forms of biases such as the use of derogatory words and expressions, stereotypes, and certain syntactic cues. Recent literature on automated hate speech detection is a testament to

this (Davidson et al. 2017), (Z. Zhang et al. 2018), (Cheng et al. 2021). Common NLP methods used to this end include feature-based models such as n-grams and Support Vector Machines (SVM), as well as deep learning models featuring variants of neural networks like Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). The general consensus on this application is that deep learning models tend to outperform traditional feature based models (X. Zhang et al. 2015); (Devlin et al. 2018), (Y. Chen and Pan 2022). Their more complex architecture enables them to capture more intricate, nuanced relationships in the data. When it comes to implicit forms of bias, the main testing methods for it in a machine learning model are word embedding methods and generative tasks for deep learning. While these methods are not used in this study, they are essential for understanding the context of the NLP and bias.

Bolukbasi et al. (2016) introduced one of the first papers looking into word embeddings and cosine similarity for a method to detect and de-bias word embeddings. They proposed the first method to mitigate this bias by ensuring equidistant pairs for commonly gendered words and their base pairs of pronouns. This was done by testing analogy puzzles substituting common tropes and measuring the distance in the underlying Word2Vec cosine similarity. While parts of it are still used today, current debiasing methods often use a more robust algorithm for word vectors similarity. This is more effective but prone to bias from the programmers. As Nissim et al. (2019) argues, when we asses the bias of text ”how much bias leaks in from our own views, preconceptions, and expectations?”. Another popular bias testing method is WEAT (Caliskan et al. 2017). WEAT is a methodology that utilizes IAT (implicit association test) literature by pairing embeddings of a set of paired target words and an attribute. These pairs and attribute scores are calculated to find the effect score for how biased the underlying GloVe embeddings are. This testing methodology is used extensively to this day with over 1,600 citations.

Today, embeddings have moved beyond Word2Vec and GloVe. Contextualized word embeddings are more commonplace for their ability to capture more nuanced data about a word. Zhao et al. (2019), utilizing the context from Bolukbasi et al. (2016) and Caliskan et al. (2017) found that contextualized embeddings also inherited bias from the underlying training set. Utilizing a ”bias

free" data set they were able to prove that ELMo inherits bias the same way other embedding methods do. The work on embeddings by prior researchers has lead to the development of recent papers like Bailey et al. (2022), where word embeddings were trained on a common crawl of the internet and cosine similarity was tested to find that people continues to be most similar to men.

Deep learning models are used and favored for their high accuracy and ability to be trained unsupervised. These models are more "black box" so testing methodology has changed. The most common method is through a generative framework where sentences are masked and the model is given sentences that would often lead to a gender biased responsive like "BLANK plays football". This is observed in papers like Sinha et al. (2021); Nangia et al. (2020); and Bartl et al. (2020).

Gendered language ranges from the use of third person pronouns - "he" or "she" - to using specific language, such as honorific or occupational titles - such as "Sir" or "businessman", respectively. It is an important consideration to have, as it can create the impression that certain societal and business roles are more suitable for men or women. Garg et al. (2018) prove this in their study "Word embeddings quantify 100 years of gender and ethnic stereotypes". Their findings conclude that job ads in certain professions use words that are associated with one gender more than the other, thereby discouraging members of a particular, undesired group from applying.

Lastly, another salient facet to consider is that of intersectionality in bias. Biases related to one minority group can interact with those related to others. For example, bias related to gender can mesh with biases related to race, ethnicity, or other factors in nuanced ways. As Vaidhyanathan 2018 argues, intersectional biases are at particular risk of getting amplified, especially in contexts where algorithms seek to maximize engagement, like social media platforms.

## 7.3 SIMILAR WORK

The most similar works to this paper are found in Asr et al. (2021) and Chao et al. (2022). In Asr et al. (2021) they create an automated ingestion pipeline of Canadian news sources to generate an analysis of the gap between men and women quoted in news articles. They have found that men are quoted on average 3 times more than women. Chao et al. (2022) applies a very similar technique

as we do but with the context of media bias. Utilizing tweets from news sources as their inputs and outside rating agency for labels, they use text data to infer different coordinates of left-right leaning and high-low bias. Among the models explored, they found that their Bi-LSTM RNN was the best performing one, capable of predicting bias with a high accuracy without the use of a human oversight panel like their ground truth data.

Part IV

# METHODOLOGY

# 8

# DATA DESCRIPTION

## 8.1 CORPERA

This study examines 3.3 million full-text BBC articles over the years 2010-2022. The data was retrieved in comma-separated values (csv) and JavaScript object notation (JSON) data-exchange formats and its total magnitude was 16.75 gigabytes. Each row contains a full-text article feature, and 8 additional columns of metadata. These include: author, date, URL, title, short description, category, BBC article number, and date of scraping. The dataset features a variety of qualitative data. Most features, due to their category-oriented nature, are of a nominal qualitative data type: author, URL, category, article date, and scraping date. The rest - title, short description, and the full-text articles - are of textual data type as they consist of unstructured, natural language text data.

### 8.1.1 *Data Collection*

The provenance of the data used in this study is the BBC Online site, however, it was collected by a third party. Over the past year, the paywalls in media establishments have become more sophisticated, and the granting of permissioned access to text data more difficult to obtain. This is despite the academically-oriented nature of this research. Therefore, the data was purchased from WebCrawlers - a site specialized in web crawling and scraping articles from different media sources including the BBC, the New York Times, Huffington Post, and Al Jazeera, among others. An

investment of 2180 Danish kroner was necessary for this end, and BBC data was selected because it was the most voluminous out of all other sources.

### 8.1.2 *Data Quality*

High-quality data is key to generating valuable data insights. Based on the data quality framework brought forth by (Cai and Zhu 2015), this data can be categorized as high-quality, as it meets all of the five dimensions of data quality: usability, reliability, completeness, consistency, and relevance. It is highly usable because it originates from a credible source: a verifiable, reputable journalistic organization. It is reliable as the data contains the original state of the source information. The data is complete as there were no missing values nor duplicates in the features of interest for this study, namely, author, data, and full-article text, and it is consistent because the inputs were highly uniform. The data, in addition, is highly relevant to the theme of the research question, and its formatting and contents were understandable.

### 8.1.3 *Natural Language Text Data*

The features of a textual data type in the BBC dataset exhibit high dimensionality and sparsity. Given a text corpus, it is likely that the vocabulary will be varied, featuring a combination of common and uncommon words (Drikvandi and Lawal 2023). It is also likely there will be sparse occurrences of a specific word or sentence structure due to the variable nature of natural language text (Drikvandi and Lawal 2023). This is not a limitation but rather a reality that must be acknowledged when working with this type of data.

## 8.2 LEXICONS

In the development of the methodology for testing the model's performance and level of bias, outside lexicons were referenced and utilized to bring insights and context to the models and their performance. Four different lexicons were selected based on their relevance in the context of the patriarchy and the development process of the lexicons. Emphasis was placed on datasets that were

developed more recently and ones that had a human processing element in an effort to capture more semantic substructures and complex concepts.

### 8.2.1 *Violence*

Calling back to the systems of oppression 5, The Media as a system must perpetuate the system. One of the methods of oppression to perpetuate is through violence. The models can be examined for trends and prevalence of violence words. In this paper The Grievance Dictionary (Vegt et al. 2021) was utilized to develop a violence lexicon. The Grievance Dictionary utilized threat experts to generate categories of words, humans to propose terminology, machine learning to expand the lists, human annotation for the suitability of words and the magnitude of the word, and a final human overview. This rigorous process led to the formulation of over 24,000 categorized and rated words. For the development of the violence lexicon words from the 'violence' category were used. Because the training data for the study utilizes single word tokens, multi-word terms were removed from the lexicon. See Appendix 18.1.6 for a full list of words.

### 8.2.2 *Power*

In 2017 Sap et al. 2017 quantified a relationship between power, agency and the theme and agent in a sentence. They utilized connotation frames as the basis for developing this relationship to try and show how words can take and remove power. To develop these scaled lists they used AMT crowd-sourcing to rate 1,700 transitive verbs on the power differential in the text. An important distinction in this paper, and why it was used, is the delineation that a word does not linearly have more power than another word. Sap takes the perspective that power is a scale between the agent and theme in the sentence. Examples are provided in table 3 highlight the flow of agency and power in sample sentences.

For our lexicon we utilized the power words that have greater power to the agent in the sentence. These are words that can be seen as "high power" to the speaker in the sentence, which with our classification method, we assume belongs to the gender of the classified sentence. This is supported

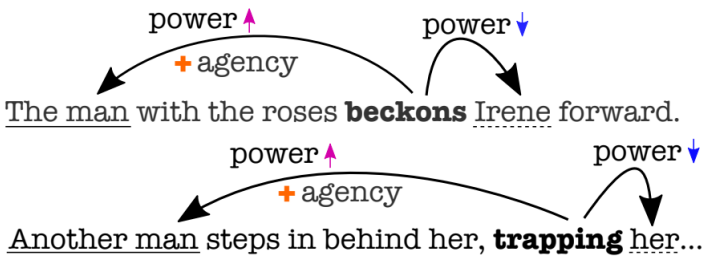


Figure 3: Power, Agency Theme & Agent in Sentences from Sap et al. (2017)

by the theories of the patriarchy 5.2 where men as a class are valued for having more power at the expense of women. This idea is supported by the systems of oppression, including The Media. The entire list of words can be seen in Appendix 18.1.3. Agency words from this paper were not used as the list was less comprehensive and less recent than Lawson et al. 2022.

### 8.2.3 *Agency/Communality*

For agency and communality words the paper by Lawson et al. 2022 was used. The process of generating their word list was a literature review, refinement, and then a rating system by crowd-sourced workers to rate the words. Importantly for this paper agency was defined as: *"An individual's striving to be independent, control one's environment, and to assert, protect, and expand one's self. Agentic individuals are autonomous and individualistic. They strive to achieve their goals, experience achievement, and master their environment, even if they have to conquer obstacles and dominate others. Agency-oriented individuals experience fulfillment through their individual accomplishments and their sense of independence and separateness from others."* (Lawson et al. 2022, SI Appendix, pg. 7 ).

Communality was defined as: *A person's striving to be part of a community, establish close relationships and connect with others. These individuals are empathetic and understanding. They strive to closely relate and cooperate and merge with others, even if they sometimes must sacrifice their individual needs for the common good. They experience fulfillment through their group accomplishments, close relationships, and a sense of belonging."* (Lawson et al. 2022, SI Appendix, pg. 8 ).

These definitions align with the expectations of behavior outlined by the patriarchy where women are more communal through an assumption of empathy and care-taking and men are more agentic through their position as the dominant class in society, allowing them to control their environment and expand themselves. See Appendix 18.1.4 for a full word list

### 8.2.4 *Appearance*

For the appearance lexicon, the words from Garg et al. 2018 were used. These words were developed by aggregating online lists to form the basis for comparing word embeddings to highlight bias. While the methodology for forming this list is less than ideal, this list is used repeatedly in other studies like Chaloner and Maldonado 2019 and Kozlowski et al. 2019. To see the full list see Appendix 18.1.5.

## 8.3 GENDER CLASSIFICATION

The models used for this thesis are intentionally supervised methods. While restricting to a supervised method reduces the overall accuracy of the classifier, it helps to keep the model's performance contained for the research question. For more information about model selection see Section 12.2.2. To generate the list of words for classification many aspects were taken into consideration: What constitutes a gendered sentence?; What is indicative gendered language?; What terminology will be considered?.

A gendered sentence in this study was considered one where the gender present in the sentence is binary. For more clarification on this process see the Preprocessing Section 11. Gendered indicative language was then divided into subcategories of types of gender indication language. A framework from (Bigler and Leaper 2015) was used to define gendered roles and words but was also expanded on to include: pronouns/identity, honorific titles, familial roles, names, regency honorific titles, gendered occupational titles. Most papers that use classification or other supervised methods to look for gendered bias use a different category or blend depending on the task. Bolukbasi et al. 2016, often considered one of the most influential studies on NLP and bias, utilized a limited list

| Pronouns/Identity | Honorific Titles | Familial Roles | Names | ~~Regency Honorific Titles~~ | ~~Gendered Roles~~ |
|---|---|---|---|---|---|
| he -- she | Miss -- Mister | Mother -- Father | John -- Linda | ~~Baron -- Baroness~~ | ~~Witch -- Wizard~~ |
| his -- her | Mrs. -- Mr. | Grandmother -- Grandfather | Adele -- Boris | ~~King -- Queen~~ | ~~Actor -- Actress~~ |
| girl -- boy | Sir -- Lady | Stepson -- Stepdaughter | Danielle -- Daniel | ~~Lord -- Lady~~ | ~~Waiter -- Waitress~~ |

Figure 4: Table of Classification Words

of pronoun and identity words. While this approach is simple, it neglects the underlying structure of news data. Many sentences in news data will use a descriptor of a role to describe the person in subsequent text. See Figure 4 for a concrete example of what words were used and what kinds were not.

This tendency led to the expansion of classification methods by utilizing names and familial roles. While regency honorific titles work well with British news data, it has the drawback of being too encompassing with the tendency to use ”king”, ”queen”, and ”princess” as nicknames and pejorative terms. Gendered occupational titles were also selected for exclusion due to lack of similar studies including them and the proliferation of ”male-default” language leading to over classification. For a full list of selected terms see Appendix 18.2

# 9

# EXPLORATORY DATA ANALYSIS

Prior to developing an approach to answer the RQ, some exploratory data analysis (EDA) was performed to understand the data. To retrieve meaningful insights, full-text articles are split up into sentences. This is done to facilitate the identification of whether a piece of text concerned males or females. Said metric is used as a proxy for the volume of representation for each gender.

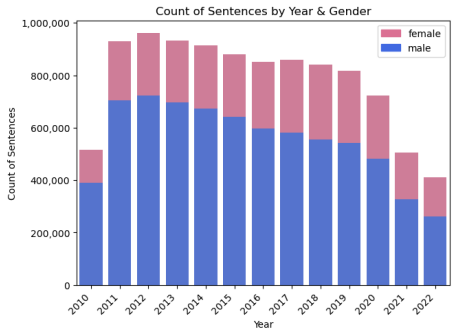


Figure 5: Proportion of Male and Female Sentence in 2010-22

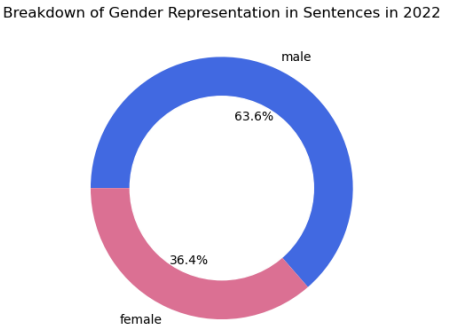


Figure 6: Breakdown of Male to Female Sentences in 2022

The first important insight revealed by the EDA is that females are under-represented vis-à-vis males in every year studied. As seen in Figure 5, there is a consistently large difference in the proportion of sentences where the subject is male or female that is sustained over all the years studied. Said disproportion ranges from 23% to 36% in 2010 and 2022, respectively. Overall, the average proportion of sentences where the subject is female during the 12-year span is only 28%. Over time, however, there is a consistent downward trend in the proportion of sentences where the

subject is male, as shown in Figure 5. This indicates that women start taking more of the space in news text as time goes by. The downward trend in the proportion of male-focused data does not translate into equal representation at any point. Notably, in 2022 only 36.4% of the news data pertains to women as seen in Figure 6. This is significant because 2022 has the highest proportion of female-focused sentences over the period studied. The key takeaway from this is that there is class imbalance between the male and female classes. Despite this, a class balancing algorithm will not be utilized. The reason for this being that the study seeks to provide interpretable results that accurately reflect the true distribution of news text when it comes to gender. The unequal nature of the proportion of female to male-focused sentences is simply part of reality.

A second interesting insight emerging from the EDA is that the distribution of news data per year is also not equal. Some years like 2012 had a significant larger volume of news, while others like 2022 had a relatively lower volume. This can be observed in Figure 5, where the total number of sentences for the former is 961,239 and 410,785 for the latter. While by no means drastically imbalanced, it is to be expected that the yearly models will perform better in years where there is more data. This is because having more data available for training enables models to better capture trends and relationships between the feature and target variables.

Semantically, a third reveal from the EDA is that news text tends to only concern one gender at a time at a sentence-level. Only 4.75% of the data is lost when removing sentences that contain gender identifiers such as pronouns, honorific titles, first names, and select gendered nouns. Importantly, this confirms that this study's approach to sentence class labelling is viable because it preserves most of the variance in the data.

To add further dimension to the data, a simple sentiment analysis exercise was performed. The Textblob library in Python was chosen to this end. A lexicon-based approach was desired to leverage its predefined rules which determine sentiment given a piece of text in an efficient, robust, and highly interpretable manner. In terms of sentiment, Figure 7 shows that sentences about females exhibit higher subjectivity and polarity than those about males. In general, this means that females tend to be more positively, but also more subjectively, portrayed. In comparison, males tend to

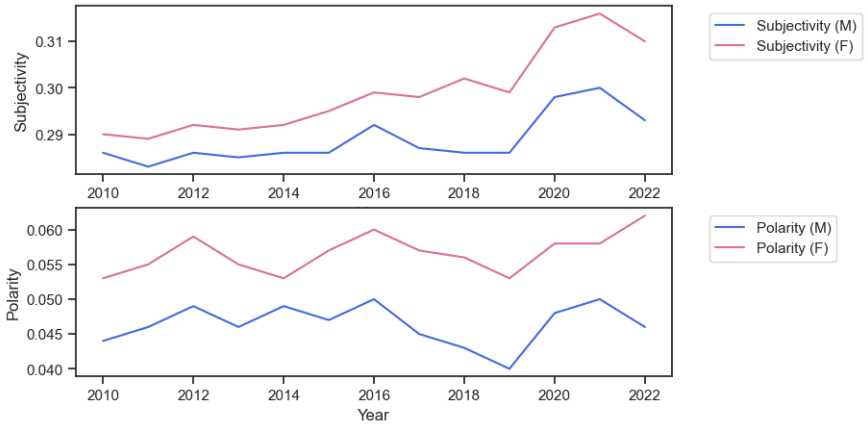


Figure 7: Proportion of Female Sentences 2010-22

be more negatively and objectively represented. This is interesting because it reveals that there is a stronger opinion-driven component in news text concerning women compared to men. Figure 7 also reveals that text concerning both genders exhibit similar subjectivity and polarity patterns. This indicates that both classes have similar sentiment characteristics over time.

From a results perspective, it would be interesting to see if the sentiment trend for male and female subjects is driven by stereotypical gender portrayals in the news text, or rather, more so by events occurring in politics and pop culture. One could hypothesize that the consistently higher polarity in female sentences suggests that women are positively portrayed because they are being represented stereotypically. Based on Figure 7, and the subjectivity trend, one could also hypothesize that female-focused sentences will likely have a stronger association with opinion words and adjectives as well.

Lastly, the EDA affirms some of the pre-processing that needs to be carried out. The most frequently featured words in yearly datasets include stopwords such as ”in”, ”the”, and ”and”. These commonly-used words are insignificant insofar as inferring meaning and take up space. Therefore, they will need to be removed. This will reduce dimensionality and enable the study to focus on features in the news text which contribute to meaning.

# 10

# METHODOLOGICAL APPROACH

This thesis approaches the research question outlined in Section 3.3 as a two-tier text classification problem. Utilizing the hypotheses developed, it examines gender bias in BBC news text at the word- and sentence levels. This approach identifies the presence of gender bias in written text using the definition and criteria of such outlined in the Conceptual Framework (4.1). It analyzes the portrayal of subjects in sentences referencing lexicons from feminist linguistics to capture gender bias based on power, agency, communality, appearance, and representation trends in the text. These five dimensions are treated as indicators of bias, and can be used to draw conclusions on the under- and mis-representation of individuals. The methodological approach consists of three steps: preprocessing, modelling, and the analysis of performance and results.

Feature extraction occurs after preprocessing and prior to modelling. The techniques used for both algorithms are different. For logistic regression (LR), *TF-IDF Vectorizer* is employed. Meanwhile, for the bidirectional recurrent neural network with LSTM layers (Bi-LSTM RNN), *Text Vectorization* is the extraction technique used. Distinct feature extraction techniques are employed because the two algorithms serve a different purpose. For instance, TF-IDF is useful for the LR algorithm because it makes model coefficients more interpretable. This facilitates the identification of the most important features for the classification task, which is important to explore *Hypothesis 1* through LR. Per contra, TF-IDF was not employed in the Bi-LSTM RNN because TF-IDF scores are solely based on the frequency of words, meaning that the relationship between words in the

sentence would not be taken into account. Syntactic choices, though, are highly relevant insofar as addressing *Hypothesis 2* with the Bi-LSTM RNN.

This study leverages traditional and deep learning machine learning techniques to extract insights that answer the different hypotheses. The traditional approaches uses LR, while the deep learning approach uses a Bi-LSTM RNN. A yearly model is run for each year between 2010-2022 using both algorithms. In total, 24 models are used to capture the insights required to address the research question. LR models are leveraged for their ability to extract coefficients. All coefficients for each year are analyzed to determine whether diction differ when the subject of a sentence is male or female, thus addressing *Hypothesis 1*. The Bi-LSTM RNN is used for its memory component, enabling it to capture the sequential elements of a sentences. It takes the LR models a step further by assessing whether syntactic choices differ in sentences where the subject is male or female. This is done through an original testing methodology that speaks to *Hypothesis 2*. Both algorithms are employed to identify whether stereotypical language, as described above, is exhibited to refer to individuals of different genders.

The aspect of volume of representation is evaluated using simple arithmetic methods. Concretely, the number of male and female sentences are quantified each year, and the insight will emerge by dividing the total count for each gender over the total counts of sentences every year. To avoid data pollution, only sentences that contain linguistic indicators of one gender will be considered. However, this is not an issue, because as learned in 9, on average 95.25% of the data each year consists of sentences that meet this condition. Traditional machine learning performance evaluation metrics are evaluated for all models to extract macro and micro insights. The metrics of particular interest in this study are accuracy and F1-score. The juxtaposition of performance between the two classes - males and females - is of particular interest when discussing the presence of gender bias as per *Hypothesis 3*.

# 11

# PREPROCESSING

In NLP the preprocessing is the process to try and turn language into encodable text and help to limit the null space when generating a function. Every preprocessing pipeline is different because every machine learning technique, language represented, and question researched requires a different kind of preprocessing to allow the data to best fit the task. For this thesis, the sentences needed to be classified as male, female, or none, separated into a list of words, and language that could introduce bias into the model was removed.

The preprocessing pipeline begins by taking every news article and separating it by sentence using simple punctuation delimiters like [!,?..]. This is a fair assumption for our dataset as all of the news articles have been published and utilized a human editor. Each sentence was then run through some simple regex to remove HTML errors, errant and unnecessary punctuation, and any email addresses or URLs. Every word in the sentence was then tokenized (a process to change a string of words into a list of words) and every word with only one character was removed. By removing small words we are able to iterate faster through the tokens and the single character data is largely irrelevant for the research question.

One of the most complex parts of processing language is balancing speed and accuracy for machine learning. While models perform best with extremely large amounts of data, processing power is not unlimited. This leads to decisions for how to handle words like “hope”, “hopeful”, and “hopefully”. While the base word of hope is the same, the words are a noun/verb, adjective, and adverb respectively. When asking a question about the language we use when we talk about groups

of people it's important to realize that each part of speech helps to indicate who the information in the sentence is about but if we leave all of the parts of speech fully intact, our corpora size increases substantially. For this thesis, all verb tenses were changed to be present tense but all other forms were kept as is. This choice was made deliberately to allow for a distinction between adjectives and nouns. When examining the word choice about groups, losing the ability to distinguish a descriptive adjective versus a noun, loses a significant amount of detail about the subject.

To lemmatize these verbs a part of speech tagger was utilized to detect the POS for every word in the text and words that were all kinds of verbs were reverted to the present tense. At the same time, the proper noun tags were saved to a separate list to detect and classify the gender of the names mentioned in the function. For each sentence that contained a proper noun, the sentence was transformed into a syntactic tree to detect if the word was a name vs a place. The names that were then called with Genderize.io (utilizing the same process as (Asr et al. 2021)) to detect the gender and calls were cached.

| Sentence Examples | Output Text | Gender Count | Classification |
|---|---|---|---|
| "On June 1st, Isabel was seen outside her apartment wearing Balenciaga shorts and a red halter top" | ['see', 'outside', 'apartment', 'wear', 'short', 'red', halter', 'top'] | Female: 2<br>Male: 0 | Female |
| "But rather than bring a family member, a friend or even a pet, he splashed out NZ$200 (£100) on a clown called "Joe"[M], who sat making animal balloons during the meeting." | ['rather', 'bring', 'family', 'member', 'friend', 'even', 'pet', 'splash', 'clown', 'call', 'sit', 'make', 'animal', 'balloon', 'meeting'] | Female: 0<br>Male: 2 | Male |
| "Homes'[U] attorneys said she and her partner Billy[M] Evans were planning to attend a wedding and hoped she would be acquitted." | ['attorney', 'say', 'partner', 'plan', 'attend', 'wedding', 'hop', 'would', 'acquit'] | Female: 3<br>Male: 1 | None |

Figure 8: Table of Sentence Preprocessing

The preprocessed words were then compared to the Gender Classification Lexicon 18.2. These words were used as indicators of gender in a sentence. The words were counted for each sentence and only sentences with one class containing a gendered word were kept. This allows for sentences that only talk about men or women to be used for training. While in a perfect world, sentences would be kept for the corresponding class if the agent in the sentence was of that gender but this is extremely complicated to do without the usage of a large language model. Some example sentences can be seen in Figure 8

Once these counts were generated and stored all gendered words were dropped from the model to ensure that leakage was not occurring within the model. The data was then further cleaned by dropping stop words, expanding contractions, and removing words smaller than 2 characters. This process left a list of preprocessed words that can then be numerically encoded and trained with a model.

# 12

# MODELS

## 12.1 LOGISTIC REGRESSION

The first tier of the classification problem explored in this thesis utilizes a logistic regression (LR) algorithm to ascertain whether the diction choices change when a sentence concerns males or females. The output of this technique addresses *Hypothesis 1* - through word coefficients - and its performance, *Hypothesis 3*.

Logistic regression (LR) is a traditional model in machine learning. Its basic nature may make it seem unsophisticated but it is a highly effective algorithm commonly used for text classification tasks in NLP. It is easy to train and understand, and as a matter of fact, many classification problems innately have a logistic distribution (Shah et al. 2020). In text classification applications, NLP literature is consistent that LR performs better than other baseline machine learning classifiers. Examples of such include decision trees, K-nearest neighbors, naïves Bayes, random forest, and support vector machine (SVM) (Pranckevičius and Marcinkevičius 2017), (Wang et al. 2020), (Shah et al. 2020), (Wendland et al. 2021), (Hassan et al. 2022).

In this study, 12 different LR models are run. One for each year in the 2010 to 2022 period. The same algorithm is used yearly, however, each year is considered to be its own model given that hyperparameters change due to GridSearchCV. The LR models were run using a nested function which takes in a year, the dataframe, a label column, and finally, a text column. The components of the nested logistic regression function are each outlined in the sub-sections immediately below.

All aspects of the modelling were carried out using Python's Scikit-Learn library (Sklearn) which specializes in machine learning and statistical modelling.

### 12.1.1 *Feature Extraction*

In the nested LR function the first thing that occurs is the definition of the predictor and response variables. These are configured to based on the label and text column arguments inputted into the function. Next, the features of the predictor variables are extracted using the TF-IDF technique, and the *TfidfVectorizer* function. The decision to combine TF-IDF and logistic regression is based on a solid body of work in NLP showing that said combination yields the highest accuracy compared to other extraction mechanism and model combinations (Pranckevičius and Marcinkevičius 2017); (Wendland et al. 2021); (Shah et al. 2020).

Alternatives for feature extraction such as word embeddings were also considered due to their popularity and prevalence in NLP. Ultimately, TF-IDF prevailed over a word embedding technique given the specific demands of this use case. First, the data handled in this study is natural language text, meaning that it is highly sparse. As such, in this application, word embeddings would not be as effective as TF-IDF as they require large amounts of data to learn the relationship among words (Dessi et al. 2021). Second, given that the regression coefficients are of interest in this study, it is worthwhile to capture information about words that could potentially not be present in the training data. TF-IDF can provide this information while word embeddings cannot (Dessi et al. 2021). Third, literature focused on how NLP techniques can amplify societal biases have pointed out that word embeddings can magnify biases that exist in the language data that they are trained of (Bolukbasi et al. 2016), (Caliskan et al. 2017), (Garg et al. 2018), and (Gonen and Goldberg 2019). This is because said technique picks up on subtle association between words and topics. Hence TF-IDF, as it is focused on word frequency more so than association, is considered a safer choice when it comes to this particular application. Finally, TF-IDF is known to be a more computationally efficient technique than word embeddings. Given the moderate computational power accessible for this study, TF-IDF also made sense in this regard.

### 12.1.2 *Data Split*

Subsequent to feature extraction, the data is split into training and test data utilizing the *train_test_split* function. The first input into said function is a numerical matrix where each row corresponds to a piece of text - a sentence - and each column corresponds to a specific word. The values in the matrix are the TF-IDF scores for each word in each document. The second input is a Pandas dataframe containing the class label for each sentence. These labels are determined by a function that was built to capture only sentences that have indicators of one gender based on pronouns, honorific titles, given names, and selected gendered nouns as previously done in Bolukbasi et al. (2016). The test size was set to 20% as the yearly datasets are relatively large, and thus, would be representative of the sample data when it comes to evaluating model performance (Rácz et al. 2021).

### 12.1.3 *GridSearchCV & Architecture*

After splitting the data, the GridSearchCV technique is applied. GridSearchCV finds the best combination of hyper-parameters for the model based on a specified hyper-parameter. Given that the data for each year is different in terms of volume and proportion of male to female sentences, the hyper-parameter grid was purposefully made ample. This space consisted of one constant - the random state - and four hyper-parameters featuring a list of different options for regularization (penalty), inverse regularization strength (C), solver, and finally the class weight.

The penalty hyper-parameter enables one to select the type of regularization applied to the model. In logistic regression, there are two types of regularization: Lasso (L1) and Ridge (L2). Both of them add a penalty term to the cost function. In the former, it is proportional to the absolute value of the coefficients, while in the latter it is proportional to the square of the coefficients (Muller and Guido 2016). Generally speaking, the first is better for feature selection while the second is better to prevent over-fitting (Muller and Guido 2016). The inverse regularization hyper-parameter $C$ handles the trade-off between good model fit on the training data and generalization. The larger $C$ is, the smaller the regularization, and therefore the more prone an algorithm is to over-fitting

the training data. The specified values for *C* tested follow a logarithmic scale: 0.1, 1, 10, and 100. Using this scale, values that are one-order magnitude apart, both small and large, can be tested (Fernandez-Martinez and Fernandez-Muniz 2020).

The random state is set equal to 42. This hyper-parameter is arbitrary and does not affect model performance. Rather, it ensures consistent, reproducible results. The random state could take any integer value and accomplish the same thing. However, 42 is commonly used as an inside joke in scientific communities in a nod to the novel *Hitchhiker's Guide to the Galaxy* where it is the answer to the ultimate question of life (Sahagian 2020). The class weight hyper-parameter tested for two values: "balanced" or 0:0.3, 1:0.7. Class weight is an important hyper-parameter especially when modelling data exhibiting class imbalance. As revealed in the EDA 9, while the data is not imbalanced, its distribution is uneven across years and proportion of sentences across gender. Therefore, including this hyper-parameter to adjust the weights inversely proportional to the class frequencies observed in the input data. All possible solvers for LR were included in the hyper-parameter grid except for lib-linear. This is because said solver is better-suited for relatively smaller datasets. Therefore, the solvers considered were "lbfgs", "newton-cg", "sag", and "saga".

The LR classifier is inputted into the *GridSearchCV* function, containing the parameter grid with the hyper-parameters outlined above, and is cross-validated using 5-fold splitting strategy. The test and train data are then fit into the *GridSearchCV* function. The classifier containing the best performing combination of hyper-parameters is ultimately utilized to generate an output based on the test data.

The final LR architectures based on the GridSearchCV hyper-parameters, varied across years. The hyper-parameters subject to most variation were the inverse regularization strength - which took on every value in the logarithmic set across time - and the class weight, which was 0: 0.3, 1: 0.7 from 2010-2016 and then balanced in years 2017-2022. The penalty, solver, and random state hyper-parameters remained constant each year. Thus, a Ridge regression using a "newton-cg" solver was ultimately run for year.

### 12.1.4 *Performance & Coefficients*

The logistic regression nested function also contained two more functionalities related to the capture of results. A classification report recording the main classification metrics - accuracy, precision, recall, and F1-score - is generated using the *classification-report* function. These metrics are useful to evaluate model performance, and to identify signs of over- and under-fitting.

The coefficients for the classifier with the optimized hyper-parameters are also extracted using the *coef* function. These coefficients are sorted to show the feature name - in this case the words - and the coefficient value. All coefficients are extracted to be able to analyze the trend of all words over time. This is key to draw conclusions between the diction choices every year.

## 12.2 BI-LSTM RNN

The second tier of the classification problem investigated in this thesis employs bidirectional recurrent neural networks (Bi-LSTM RNN) to assess syntactic patterns and bias exhibited in model performance. The performance aspect of this algorithm talks to *Hypothesis 3* while a testing methodology created for this model evinces *Hypothesis 2*. A Bi-LSTM RNN is an adequate choice of model for the latter hypothesis as it concerns the syntax of a sentence. Through its memory component and hidden layers, a Bi-LSTM RNN is capable of identifying sequential trends in the natural language text. Specifically, trends which the LR models - or other traditional machine learning approaches - cannot. The incorporation of LSTM cells in a bididrectional RNN is an effective, battle-tested approach for text classification in NLP as demonstrated by Bangyal et al. (2021) and Kashid et al. (2023).

### 12.2.1 *Architecture*

The Bi-LSTM RNN architecture implemented herein is inspired by Kashid et al. (2023). Their recent paper - using a similar skeleton - yielded impressive performance results when classifying the text of 4 million Amazon reviews. However, several modifications are made to their architecture to meet the specific demands of this study. The Bi-LSTM RNNs are executed using Python's

built-in Tensorflow platform, commonly used for machine learning. The Keras and Sklearn libraries were used to handling all other aspects of modelling.

Two essential steps occur before modelling. First, the input pipeline is set up. This is done by shuffling the data for training and creating batches of text and label pairs. Shuffling, batching, and prefetching operations are performed to this end. Relevant hyper-parameters selected for this include a buffer size of 50,000 and a batch size of 64. For the former, 50,000 is a good compromise for input sequences configuration as it keeps the training fast but also accurate as TensorFlow (2021). For batch size, an integer to the power of two and within the range of 16 and 512 is traditionally used. In this case, the batch size of 64 performed best.

The second step involves defining an encoder variable which stores the encoded, raw text data. The data is encoded using a Keras *TextVectorization* layer. The only hyperparameter configured in the encoder is maximum number of tokens equal to the vocabulary size. The vocabulary size was set at 60,000 words to capture 99.8% of all the variance in the English language (W. Chen et al. 2019). This is much in line with the aforementioned preprocessing mindset in that it is quite liberal. A relatively larger vocabulary size than what is realistically needed was deemed an appropriate course of action to ensure the Bi-LSTM RNN could effectively learn to distinguish similar words with different meanings (W. Chen et al. 2019). This is, after all, an important consideration when it comes to addressing *Hypothesis 2*.

The Bi-LSTM RNN model proposed by this study consists of 7 layers, as can be observed in Figure 9. It is built using the Keras Sequential class. The first layer is the encoder variable described above, where the text is converted to a sequence of tokens. The second layer is the Embedding one. This layer takes in 3 arguments: the number of unique words in the vocabulary used in the encoder, embedding vectors with 256 dimensions, and a masking parameter which adds special padding token for sequences represented by value 0. The output of this layer converts each input token into a dense vector representation. This representation - the dark blue layer in Figure 9 - captures the semantic meaning of each token, enabling the model to learn complex relationships between the different tokens in the input text.

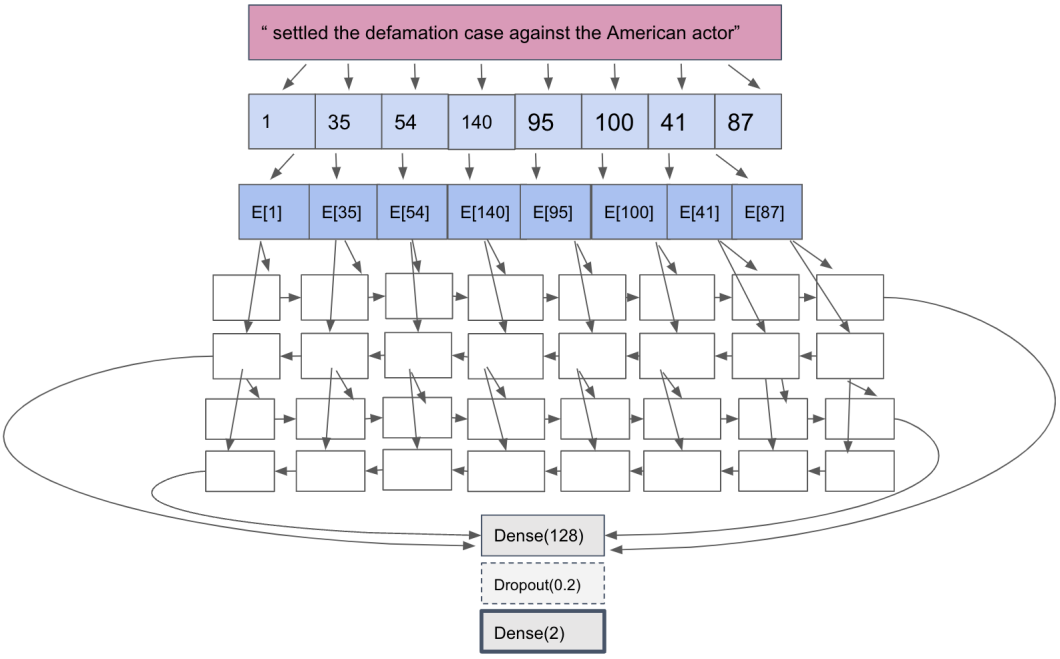


Figure 9: Bi-LSTM RNN Architecture Using a Sample Sentence

Two Bi-directional layers are subsequently stacked on top of the Embedding layer. This deviates from Kashid et al. (2023) who proposed only one bidirectional layer in their Bi-LSTM RNN architecture. Two bidirectional layers are employed instead of one to better capture syntactic information, as it is relevant to addressing *Hypothesis 2*. In this manner, the input sequence gets processed in both forward and backward directions twice. These layers are not only bidirectional but they are also long-short term memory-based (LSTM), with 128 and 64 hidden units for the first and second bidirectional layers, respectively. LSTM-based cells were incorporated as they tend to outperform traditional neural networks applied to sequential tasks due to their memory components (Yao et al. 2019). This means, that on top of understanding complex patterns through its hidden layers, it is also able to capture context. The first bidirectional LSTM layer is configured to return a full sequence of outputs for each input while the second only returns a single vector summarizing the information from all the input sequence.

The output layers of the Bi-LSTM RNN include 2 dense layers, and a dropout layer in between. These layers are shown in varying hues of grey in Figure 9. The first dense layer has 128 neurons and applies a Rectified Linear Unit (ReLU) activation function. ReLU is mathematical function

that linearly transforms the input, returning the input value if positive and zero otherwise. ReLU is utilized due to its ability to learn nonlinear relationships between the input and output, and for capabilities in learning sparse representations of input data (Xu et al. 2015). Therefore, it is a good pick from NLP applications. ReLU is an effective activation function also for the way it relieves the vanishing gradient problem that occurs as a result of backpropagation (Olimov et al. 2021). Kashid et al. (2023) also use ReLU in the dense layer to avoid overfitting. The dropout layer, sandwiched between the two dense layers, has a dropout rate of 20%. Meaning, the model randomly drops 20% of the neurons in this layer. This regularization technique is implemented to enhance the generalizability of the Bi-LSTM RNN and to prevent overfitting.

The final layer of the Bi-RNN LSTM is a dense layer, containing 2 neurons. This is another deviation from Kashid et al. (2023). The neurons correspond to the number of output classes possible in the classification task. The output of this layer is vector of length 2, which captures the Bi-LSTM RNN's predicted score for both the positive and negative class. The predicted class for a given input is the one with the highest. The default, linear activation function is applied in this layer. There are multiple reasons to use a linear transformation instead of a sigmoidal one. The former is better to capture a wider range of input-output relationships, ensure stability when input values are extreme, and better outlier handling (Xu et al. 2015).

The Bi-LSTM RNN model consists of a total of 14,926,722 trainable parameters. This model is compiled using the Sparse Categorical Crossentropy loss function. This seemed most appropriate given that the target variable is an integer - 0 for male and 1 for female, and that a predicted score for two classes is desired. The softmax function is called internally, by configuring the loss function to return logits instead of probabilities in the compile function. The Adam optimizer - a highly popular optimization algorithm - adapts the learning rate for each parameter during training by a step size of 0.0001. Furthermore, accuracy is used as the metric to evaluate model performance during the model training. Using this metric is standard practice in machine learning when compiling models.

The model is then fit using the model object. The Bi-LSTM RNN is fit to the training dataset for a maximum of eight epochs. The training process uses the early stopping criteria so that model

training stops if the validation loss does not improve for five epochs. The model's performance is evaluated on a test dataset previously developed from feature and target data tensors. This approach is computationally efficient and prevents overfitting the model on the training data, thus enhancing its generalizability when it comes to unseen data. The Bi-LSTM RNN is fit to the data yearly. This means that in total, twelve of these models are run. Their results are saved for ensuing performance analysis and to deploy the Bi-LSTM RNN testing methodology outlined below.

### 12.2.2 *Testing Methodology*

In order to analyze the data in a contextualized manner, to address *Hypothesis 2*, more sophisticated models were considered. While a large language model is more accurate and can provide more comprehensive insights, the model is semi to unsupervised. This lack of supervision could allow the model to pick up on cues within the text to influence the predictions for male or female. While fine tuning with supervision can solve this problem, it would require tuning with unbiased and balanced sentences. The semi-supervised step would correct the majority of the bias but understanding the codification of bias relies on understanding where the bias is to begin with. This chicken/egg problem of bias generation in models makes a supervised method a more appropriate for the task.

The RNN performed well in the classification task but because the RNN can only provide that classification weights for an input that looks like what was fed to the model the model can not provide insights with just the classification weight of the words like logistic regression. In an attempt to derive more insights from the model an experimental scoring technique is proposed. For each word that is being tested, 75 randomly retrieved unseen (by the model) sentences were tagged and a proper part of speech word substituted. This resulted in for a word list of 740 words and 75 sentences 55,500 test sentences. These were fed into the pretrained models along with the original sentence, used as a benchmark. The result was then 55,575 sentences with a weight for the positive and negative class.

To aggregate this score the Difference Score is proposed in the equation below 12.2.2. The difference score takes the difference of the baseline sentence and the calculated sentence for the same word and finds the average for the positive and negative class separately. Then the negative class is subtracted from the positive to help show the magnitude of difference in the prediction weight the model is assigning. Because this is an experimental equation testing was done to attempt to ensure the validity of the method.

$$\text{p\_class\_score} = \frac{1}{N}\sum_{i=1}^{N}(p\_class\_baseline - p\_class\_scores(i))$$

$$\text{n\_class\_score} = \frac{1}{N}\sum_{i=1}^{N}(n\_class\_baseline - n\_class\_scores(i))$$

$$\text{Difference\_Score} = \text{p\_class\_Score} - \text{n\_class\_Score}$$

To ensure the validity of the experimental scoring technique the following methods were used for methodological testing:

**1. Comparisons of similar words:**

For multiple given years, word performance was compared for word pairs that are contextually similar and semantically similar. See Table 12.2.2 for the full list of words. The difference of the difference score for the pairs were averaged for the time period and compared. The words were also compared to the performance of the average for the entire time period. For this test, the tested words were on average less than a weight of .02 away from each other. . This holds that the words are behaving as expected as the RNN captures words order and is therefore more sensitive semantic differences than the logistic regression.

**2. Comparing the differential in scores for different words in the same sentence:**

For all 75 sample sentences the average score, standard deviation, and variance were considered to compare how similarly the sentences score. The sentences had an average standard deviation of 2.2 and a variance of 8.22. These scores indicate that for each sentence the weights are dispersed widely and do not follow any directly observable patterns. This shows that while some sentences

| Word Pairs | RNN |
|---|---|
| adversary-enemy | -0.149032494 |
| assault-attack | 0.041457366 |
| bleed-blood | 0.030957861 |
| cry-scream | -0.012941956 |
| hurt-injure | 0.032035253 |
| rape-rapist | 0.044152227 |
| scrape-cut | -0.092324908 |
| trust-trusting | -0.083184694 |
| Average | -0.023610168 |

Table 3: Word Pair Comparison Scores

with word substitutions may have similar scores, other sentences have a broad range of available output scores. This supports the usage of the experimental difference score as it shows variability based on input.

**3. Comparing the trend of difference scores over time:**

Because the RNN weights are normalized and the Logistic Regression is not, direct comparisons of performance cannot be made strictly on the score of a word, but the directionality can be considered. For all words in the testing dataset, graphs were generated and directionality was compared with differently scaled y axis. It was found that words that were very stereotypical and words that were more neutral the lines were often different but words that were more ”slightly” stereotypical the lines had almost perfect tracking. This is highlighted in Figures 10-12. Since the RNN is more accurate and better at capturing subtlety, this aligns with expectations.

Due to the positive outcome of this testing methodology, the results will be presented in Section 13.2.2. They are presented as exploratory due to their experimental nature. There are quality insights from the data, but the inability to truly estimate the confidence level in the results leads to the presentation of the results with a caveat of their experimental nature and a smaller exploration.

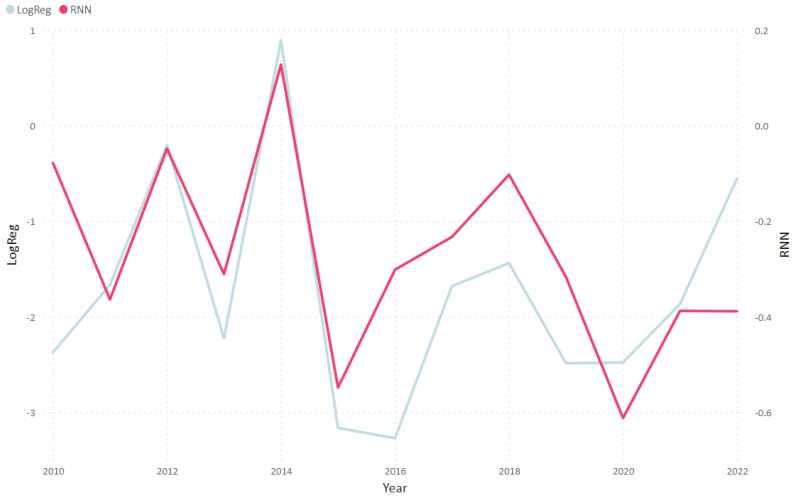


Figure 10: RNN vs Log Reg for ”Accomplice”

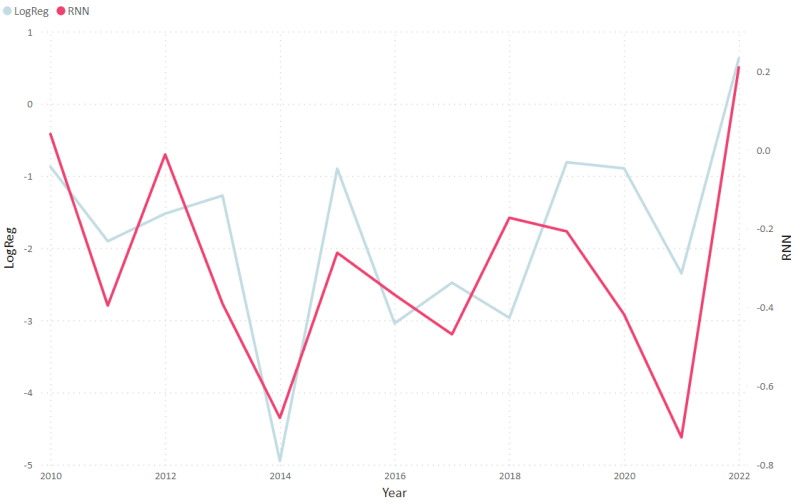


Figure 11: RNN vs Log Reg for ”Bald”

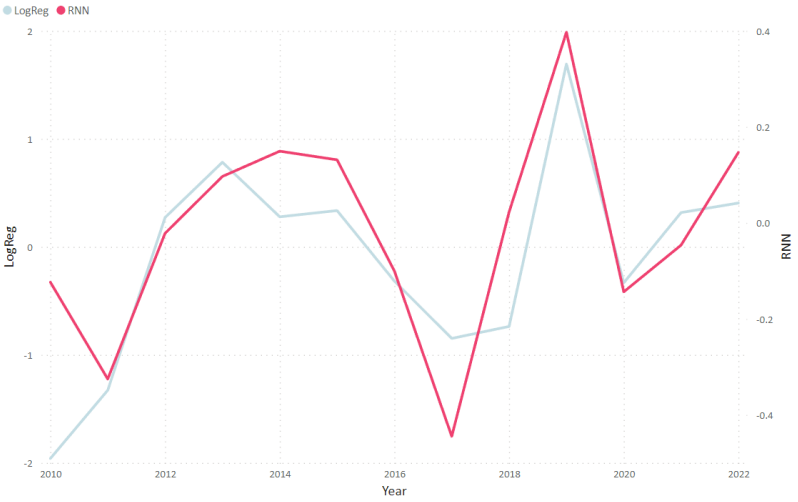


Figure 12: Bi-LSTM RNN vs Log Reg for ”Persistent”

Part V

# RESULTS & DISCUSSION

# 13

# PERFORMANCE OF MODELS

Evaluating model performance is important to discern whether they have any predictive power. The quality of the models' predictions is key to investigating *Hypothesis 3*, which states that bias is exhibited and explainable. Performance is evaluated on a yearly basis using macro aggregation, and micro-aggregation results to contextualize. A macro aggregation method is preferred for this use case because it gives each of the classes an equal importance. This is relevant given the relative disproportion between male and female sentences each year.

Multiple metrics are considered to ensure a comprehensive assessment of model performance. Consistent with the NLP literature on text classification, this study uses accuracy and F1-score as the evaluation metrics of focus (S. Liang et al. 2020); (Ahmed et al. 2022). The macro aggregation methods calculate the average accuracy and F1-scores for each class, respectively. Their micro counterparts capture overall accuracy and F1-scores, treating each instance as an individual classification task. The equations for all metrics can be found in Appendix 18.5.3.

## 13.1 LOGISTIC REGRESSION PERFORMANCE METRICS

### 13.1.1 *Accuracy*

The LR models perform well, suggesting they have predictive power. Their macro accuracy ranges from 63% to 67% over the years studied as per Table 4. The narrow accuracy range indicates that the LR algorithm can make a similar proportion of correct predictions for each year in the data. In fact, between 2013 to 2016 the macro accuracy is the same, suggesting that the text inputs during

those years do not vary significantly enough to impact the model's ability to predict the class of the sentences. In other words, the language between those years can be said to be indistinguishable.

An interesting insight from Table 4 is that in the earlier years studied the micro accuracy is bigger than the macro one. This suggests that the model performs poorly on the minority class - in this case, the female one. Posterior to 2018, this pattern is reversed, suggesting that higher accuracy is achieved on the minority class. This is seemingly driven by a higher proportion of female data. Insofar as *Hypothesis 3*, bias is not exhibited well in the LR model, as there is no obvious trend in the accuracy. However, it can be said that the bias is reducing. An all-time low macro and micro accuracy in 2022 suggest this. The observed dynamics between the two accuracy metrics over time do so as well.

| | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** | **2021** | **2022** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Macro Accuracy** | 0.67 | 0.65 | 0.65 | 0.66 | 0.66 | 0.66 | 0.66 | 0.67 | 0.67 | 0.67 | 0.67 | 0.65 | 0.63 |
| **Micro Accuracy** | 0.71 | 0.7 | 0.69 | 0.7 | 0.69 | 0.69 | 0.68 | 0.67 | 0.66 | 0.66 | 0.66 | 0.63 | 0.57 |
| **Macro F1-Score** | 0.67 | 0.66 | 0.65 | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 | 0.65 | 0.63 | 0.57 |
| **Micro F1-score (Male)** | 0.79 | 0.78 | 0.77 | 0.77 | 0.77 | 0.77 | 0.74 | 0.72 | 0.70 | 0.70 | 0.70 | 0.65 | 0.55 |
| **Micro F1-score (Female)** | 0.54 | 0.53 | 0.53 | 0.54 | 0.55 | 0.55 | 0.58 | 0.60 | 0.61 | 0.61 | 0.60 | 0.60 | 0.59 |

Table 4: Results of Yearly Logistic Regression Models

### 13.1.2 *F1-Score*

The macro F1-scores also indicate that the LR models performed well in the classification task at hand. In the period studied, this metric ranges from 57% to 67%, showcasing significant variability which could be attributed to changes in the data over time. As opposed to the macro accuracy, this metric exhibits a more obvious downward trend in the latter years studied. That suggests that the model is becoming less robust over time and could be misclassifying instances in the minority class more so than in earlier years. An explanation for this could be the relatively higher proportion of female data, perhaps portraying subjects differently than before.

Looking at the micro F1-scores, it is clear that the LR models are better at classifying instances belonging to the majority class - the male one. Inspection of the precision and recall values through time confirm that this is due to better performance in both of these metrics. 2022 is an exception to this. In this year, F1-score for the minority, female class trumps the male class by 4%. This is notable given the 25% differential between these metrics at the beginning of the time period in question. Concerning *Hypothesis 3*, the macro and micro F1-scores over time showcase that change in language is occurring. Its downward trend starting in 2018 confirms that bias is decreasing much more observably than in accuracy performance.

## 13.2 BI-LSTM RNN

### 13.2.1 *Accuracy*

The Bi-LSTM RNN models perform generally very well over time. In fact, the macro accuracy is above 70% in ten out of the twelve years analyzed, as shown in Table 5, and 75% on average for all years. This indicates that said models, on average, can correctly predict the class labels of at least 75% of the instances in the data each year. Considering that the data inputs is real-world data, this is significant, as it entails that the Bi-LSTM RNNs can effectively capture the underlying patterns in news text over the years.

| | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** | **2021** | **2022** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Macro Accuracy** | 0.86 | 0.76 | 0.77 | 0.77 | 0.78 | 0.76 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.69 | 0.62 |
| **Micro Accuracy** | 0.97 | 0.86 | 0.88 | 0.88 | 0.88 | 0.87 | 0.86 | 0.85 | 0.84 | 0.84 | 0.84 | 0.8 | 0.78 |

Table 5: Bi-LSTM RNN Micro and Macro Accuracy

From an accuracy perspective, the Bi-LSTM RNNs outperform the LR models as shown in Figure 13. The only exception is 2022. This data point reflects how neural network performance is highly sensitive to the volume of data as the combination of a relatively smaller amount of data coupled with a higher proportion of female sentences for this year compared to others, drives the macro accuracy down sharply.

In terms of *Hypothesis 3*, the Bi-LSTM RNNs exhibit a more resolute downward trend than the LR models, indicating that bias is better exhibited in it. This insight is also confirmed based on the dynamics of the macro and micro-accuracy. The latter being consistently and significantly larger than the former indicates that the model is not performing as well on the minority class, thus exhibiting notable bias. The downward trend in both metrics suggests that said bias is decreasing, however.

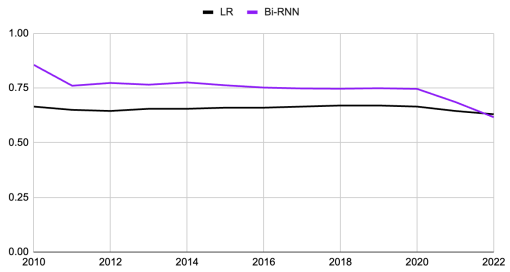


Figure 13: Macro Accuracy Comparison between Models

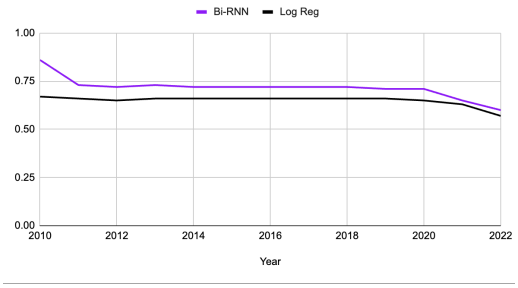


Figure 14: Macro F1-Score Comparison between Models

### 13.2.2 *F1-Score*

A lot of studies on RNNs in NLP literature are explicit in their aim to reach the best achievable F1-score Van Huynh et al. (2019) and Alsharef et al. (2022). Thus, this metric is considered determinant when assessing the Bi-LSTM models. At a high level, the models seem to be performing well on the classification task judging by the 72% average macro F1-score for 2010-2022. However, the trend shown in 14 reveals that there is a wide range in macro F1-scores over time, with a 26% difference between the macro F1-score for 2010 and 2022. This indicates that the models perform progressively worse over the period studied but also that the bias is well exhibited as hypothesized in *Hypothesis 3*.

Like the LR models, the Bi-LSTM RNNs also exhibit difficulty in correctly identifying between the male and female class. While the Bi-LSTM RNNs perform significantly better than the LR models in earlier years, their performance ends up almost converging when it comes to the F1-score. The same conclusion can be drawn from models over time: that language is changing.

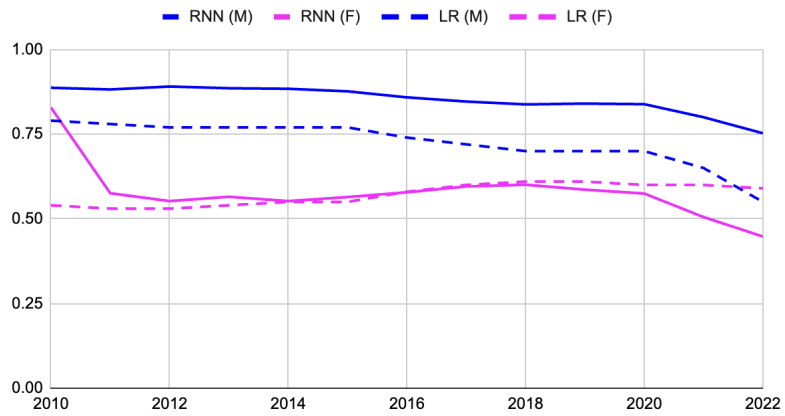


Figure 15: Class-based F1-Score Comparison in Log Reg & Bi-LSTM RNN

The comparison of F1-scores for the male and female classes reveals interesting insights about the macro F1-scores for both modelling approaches. As visualized in Figure 15, the macro F1 trend for the LR models seems to be mostly driven by the steepness of the downward curve of the male F1-score. This is so, as comparatively, the F1-score for the female class does not trend upwards as much. The same cannot be said for the Bi-LSTM RNNs. In these models, the range between the male and female class remains large throughout, becoming observably larger in the latter years of the study. The lower macro F1-score observed over time for these models can thus be said to be driven by a downward trend for both classes. This indicates that the Bi-LSTM RNNs have more trouble distinguishing between both classes, and not just the one as is the case with the LR models as time elapses. This relationship is significant when contemplating *Hypothesis 3* as it shows that bias is decreasing at a rate in which the Bi-LSTM RNNs are failing to capture some of the patterns in the text for both classes. This adaption problem shows that language is changing in a way that affects both males and females, something which reflects positive progress when it comes to bias in language.

# 14

# ANALYSIS & EVALUATION

## 14.1 INSIGHTS ON REPRESENTATION

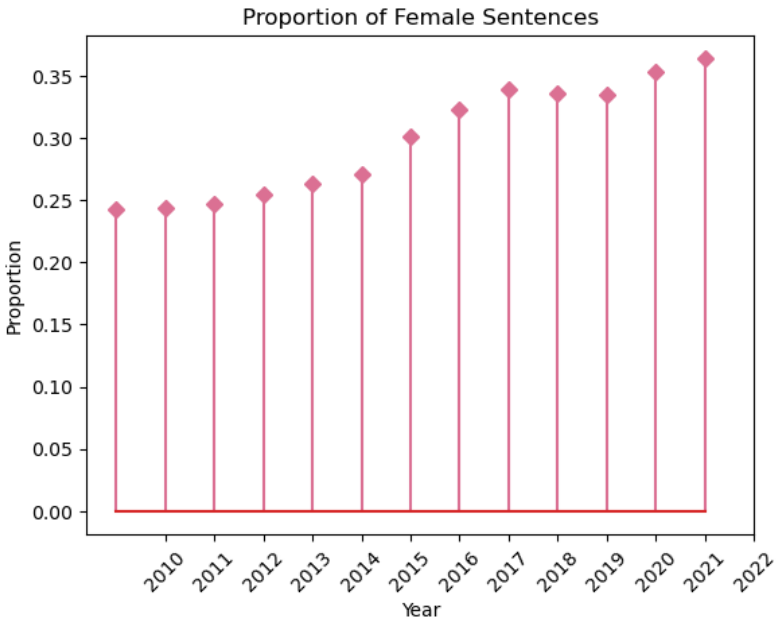


Figure 16: Proportion of Female Sentences 2010-22

Simple arithmetic operations on the yearly datasets reveal that there is an under-representation of women over time in BBC news text. This insight is the case for all years analyzed, as shown in Figure 5, where the number of male sentences each year dwarfs that of female sentences. Nonetheless, it is fair to conclude that progress has been made on this metric. Figure 16 shows a steady upward trend in the proportion of sentences where the subject is female. 2022 is the year with the highest relative representation. The ratio of female to male sentences then is 37:100. While it is not an equal ratio, it is an improvement from the first years studied. For instance, it is more

balanced than in 2010 when the ratio of female to male sentences was 25:100. The fact that the ratios are not equal at any point shows that there is bias when it comes to the gender of individuals represented. This confirms undoubtedly *Hypothesis 3*, but the trend also shows that progress has been made on this metric.

## 14.2 LOGISTIC REGRESSION

### 14.2.1 *Insights on Top & Bottom Coefficients*

In the LR function, all of the coefficients for every year are captured to assess which features in the news text play the most determinant role in predicting the class label. Several interesting insights emerge from the coefficient analysis.

First, if one were to categorize the features found in Appendix 40, it is notable that topics such as religion, reproductive health, and anatomy are most associated with the female class. For instance, ”headscarf” and ”hijab” - the first and third highest coefficient words - are tied to religion; ”pregnant” and ”breastfeed” - the second and fourth biggest coefficients - are associated with reproductive health health; and ”ovarian” and ”menstrual” - the sixth and fourteenth highest coefficients - are aspects of anatomy. These are examples among many others in 40. Notably, other words exhibiting high coefficients are related to the female appearance - in this case, their clothing. In pop culture, clothing is a stereotypical thing that women are portrayed as caring about, as it feeds into the narrative of vanity around their appearance as explained in 6.1.5. Words like ”legging”, ”bikini”, ”handbag” and ”sari” show high coefficients throughout the entire 2010-2022. This category-based gives grounds to accept *Hypothesis 3* because bias is not only exhibited but also explainable.

When conducting this analysis for the negative class, categories which would stereotypically be associated with men such as male health and leadership, are reflected in the data. This is shown by the coefficients of ”prostrate” and ”vasectomy” as to the first category; and ”knighthood” and ”strongman” as to the second. Categories such as religion and sport also seem to be highly associated with men - words like ”mujahideen” and ”papacy” have high absolute value coefficients in the negative direction, and ”batsman” and ”binman” do too. Interestingly, while religion is an

important category for both classes, in the female class they are related to religious expectations of appearance - the covering of certain body parts. By contrast, in the male class, it is related to politics. This nuanced difference, further proves *Hypothesis 1* and *Hypothesis 3* as the bias is highly evident in word choice.

A second insight from 40, is that the magnitude of the coefficients decrease over time for the top and bottom words. The direction, however, does not. Statistically what this tell us, is that the strength of the relationship between those words and the ultimate classification of a sentence based on gender, weakens over time. This is indicative that change is happening when it comes to the diction used when the subject of sentence is male or female. It is becoming more neutral overall. This aligns with what is observed in terms of model performance in 13.2.2, where the accuracy decreases and the F1-score increases over time.

#### 14.2.1.1 *Word Trends*

The overall top and bottom coefficient analysis evinces *Hypothesis 1* and *Hypothesis 3* convincingly at a high-level, while also depicting that change has occurred over time. This section is devoted to identifying whether select words that have historically been associated with women have progressed in the same fashion.

The patriarchy establishes women's role in society as being familial 6.1.5. This is excused by systems of oppression because women are unique in their biological ability to birth children. A realization of the stereotypical portrayals of women in society is displayed in BBC news as shown in Figure 17, where the coefficient for variants of the word "child" never go below 0, meaning they

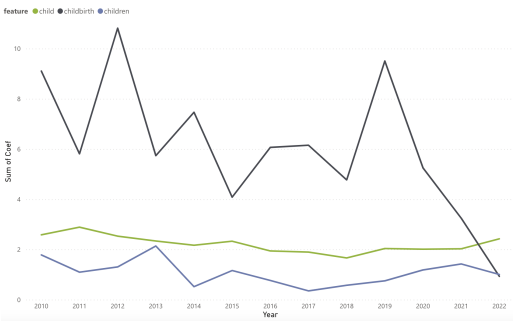


Figure 17: Variants of "Child"

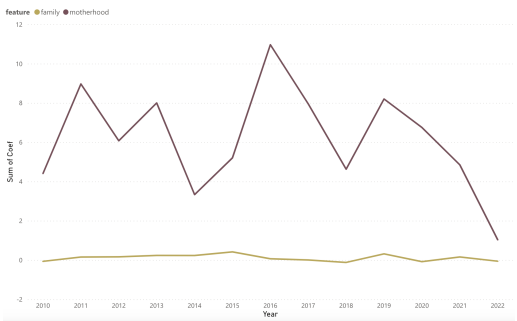


Figure 18: Family & Motherhood

are always related to the positive class. ”Child” and ”children” remain consistent in their trend over time, while ”childbirth” interestingly decreases. This could potentially be due to the fact that declining childbirth in Western countries - such as the UK - is becoming a political concern. Another compelling observation in Figure 18 is how ”motherhood” is overwhelmingly associated to women, but the term ”family” bounces around the male and female class. A potential explanation for this is that while being communal, a family is also a social institution. As feminist literature suggests, the patriarchy controls all social institutions in their position as the dominant class. Therefore, ”family” can also have strong semantic ties with the male class depending on the context.

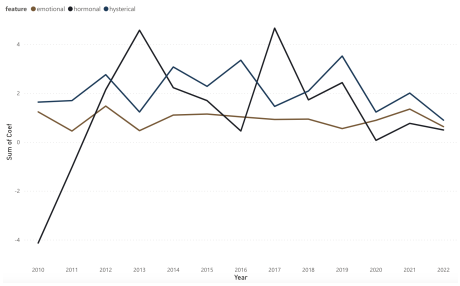


Figure 19: Stereotypical Female Emotions

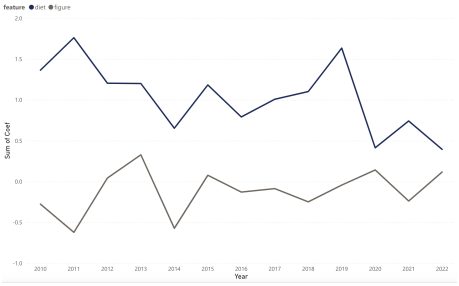


Figure 20: ”Diet” vs. ”Figure”

Another eminent topic brought up by gender studies on the matter of female misrepresentation - as outlined in 6.1.5 - is that women are often portrayed as being overly sensitive. Some despective variants in this vein include adjectives like ”hysterical”, ”hormonal”, and ”emotional”. Figure 19 shows that the coefficients for these three words are consistently of a large magnitude and positive, indicating they are associated to women. Their trends also display a similar progression, indicating they could potentially share a semantic meaning. The implication of this is that wherein the subject of a sentence is female, ”hormonal”, an a priori anatomical term, is used informally - to indicate that women are affected by their hormones to feel a certain way. The bias exhibited through this term is evident when analyzing the outlier in 2010, whereby ”hormonal” is more male-learning. What this entails is that the terms is used correctly in sentences where the subject is male, as a lot of the articles that year concerned male hormone replacement therapy. The word choice is thus serious when regarding men, and parodistic when related to women.

Gender studies highlights that women are often misrepresented as being overly concerned about their appearance. From a feminist lens, this is because of how patriarchal societies are structured, with the dominant, male class holding power, and women being objects of desire. An offshoot of this narrative is how women are commonly portrayed as ”being on a diet” or ”watching their figure”. The subtle implication is that they do so to conform to societal beauty standards, to be attractive to men. Figure 20 shows that over time the notion of dieting remains associated to the female class. This means, that women are more often represented as dieting than men, thus conforming to the gender stereotype. The word ”figure” follows a similar trend, but has lower coefficient magnitudes. In fact, it is mostly associated to men. A closer look at the articles containing said word reveals that they are mostly sports-focused. Sports, in general, is a category that is more male-leaning - proportionally and in terms of total counts. This partly explains the negative coefficients for ”figure” in 7 out of the 12 years.

#### 14.2.1.2 *Noun Pairs*

Constancy of whether word choice has evolved over time should theoretically be evinced in stereotypical pairing of nouns. One of the most common stereotypical noun pairs is that of doctor-nurse. Based on the patriarchal systems of oppression, men would be in the role of greater power - doctor - and women be more related to nurse - a role with a greater emphasis on care-taking. As Figure 21 shows, in the years analyzed, both ”nurse” and ”doctor” are more related to the female class, as their coefficients are consistently above 0. The word ”nurse” has is of a larger magnitude, as expected, while ”doctor” has a lower magnitude ranging between right above 0 and 1. Insofar as *Hypothesis 1*, it is evident that different diction is used, given that these words never change into the negatives, and barely evolve during the time studied.

Another stereotypical noun pairing is that of ”secretary” and ”manager”. Following the dynamics of power and oppression outlined in the Conceptual Framework, it is expected that the former word is more associated with the female class and the latter more so with the male. This is because ”manager” entails a certain degree of leadership, a characteristic feminists argue is more in line

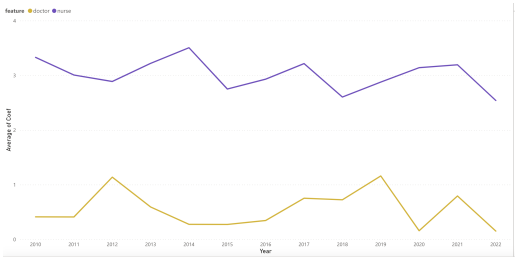

Figure 21: ”Doctor” vs. ”Nurse”

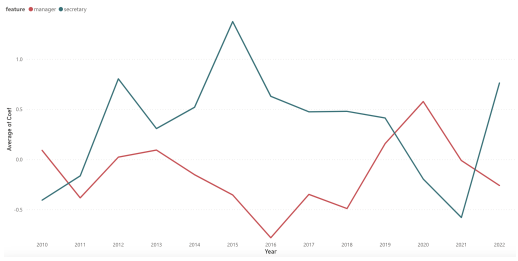

Figure 22: ”Manager” vs. ”Secretary”

with the dominant class in the patriarchy. From this same lens, ”secretary” is a much more docile word, that implies a certain degree of obedience, and thus should theoretically be more related to the minority class - women. Figure 22, it can be seen that both words flip in their relation to the male and female class, with the ”secretary” being more frequently feminine and ”manager” being more frequently masculine. Thus, in terms of *Hypothesis 1* the difference in word choice is not as clear with these two words.

Papers featuring gendered noun pairings, such as Garg et al. (2018) and Lawson et al. (2022), commonly discuss the fact that women, in the occupational realm, are portrayed as teachers. Feminist literature would argue that the patriarchy does to present women in a way that reflects stereotypical communal and nurturing attributes. Figure 23 shows that over the time studied, the stereotype holds true. The word ”teacher” never dips below 0, meaning it is consistently related to the female class - without much variance - from 2010 to 2022.

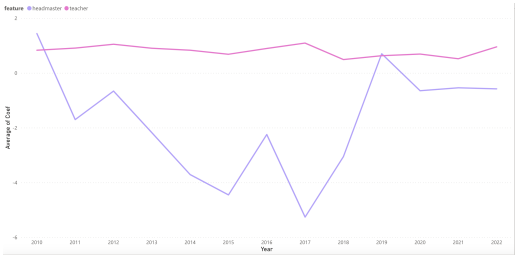

Figure 23: ”Headmaster” vs. ”Teacher”

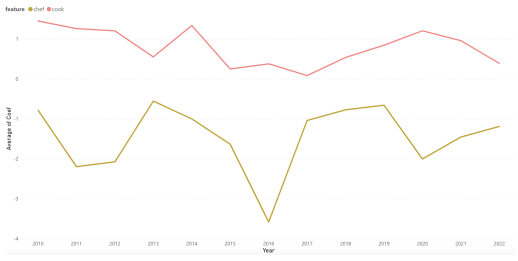

Figure 24: ”Chef” vs. ”Cook”

For comparative purposes, the word ”headmaster” was compared to ”teacher”. While also an educational role, Figure 23 shows that this word is consistently associated to men, except in 2010 and 2019 where it is feminine. The magnitude of said word in the negative direction is a lot larger

than that of teacher in the positive direction. This indicates there is a stronger relationship between the male class and headmaster, than the female class and teacher. This reveals, in the same vein as *Hypothesis 1*, that diction significantly varies when discussing individuals in positions of leadership and teaching in the educational space.

Another interesting noun pairing is that of ”cook” versus ”chef”. The word ”cook” is frequently used to describe the person cooking, and tends to connote this action within a domestic context. Whereas, ”chef” is a word to denote the action of cooking but in a more professional manner. Based on this, gender studies would suggest the former would theoretically be more feminine than the latter based on the context. Figure 24 interestingly highlights precisely this. The word ”cook” is consistently feminine, although by the end of the period studied it seems to be trending downwards towards 0. The noun ”chef”, by contrast, is always related to the male class, and follows a relatively stable trend - always in the negative coefficient zone - by the end of the period. This noun-pairing also proves *Hypothesis 1*.

In all of the noun pairs presented above, it is worth noting that the coefficient of the words that are more related to the female tend to slowly trend downwards over time, indicating they are becoming less feminine over time. However, the tendency for male-related words to become less masculine - is not as obvious. This aligns with the fact that bias is declining over time - particularly due to the movement of female-related coefficients. While bias is declining, it is still there, and the movement of the male-related noun pairings are very much a reflection of this from the viewpoint of *Hypothesis 1*.

### 14.2.2 *Linguistic Analysis*

#### 14.2.2.1 *Appearance*

Women are often portrayed as being overly preoccupied with their appearance according to gender studies research. This preoccupation with appearance is often attributed to the societal view that a woman’s primary value is in her appearance. This is a misrepresentation of females, stemming from systemic bias against them, as there is much more complexity to their persona than their

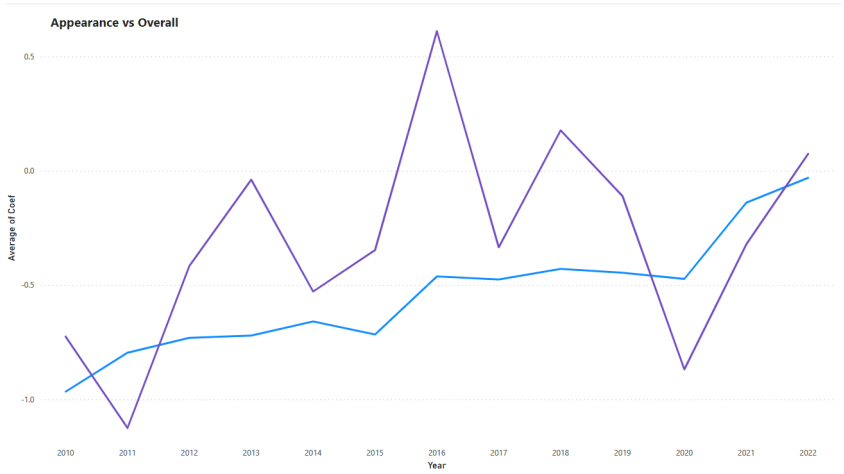


Figure 25: Trend of Appearance Words vs the Overall Trend

appearance. This study utilizes a comprehensive list of common appearance words outlined in Appendix 18.1.5, from (Garg et al. 2018).

The overall trend of appearance words shown in purple in Figure 25 exhibits considerable variation over time. In fact, coefficients range from -1.5 to 0.5 between 2010-22 which is a relatively small range. Most years, these coefficients are around 0. This indicates that appearance words are not totally masculine or feminine. This result is in line with the insights derived in the exploratory data analysis (9), which revealed that across all years, the subjectivity scores for men and women are relatively close. Subjectivity is an important metric to triangulate with when it comes to appearance, as it indicates the degree to which a text pertaining to either gender is opinion-based. This is relevant because all appearance words are adjectives, which linguistically-speaking, are more opinion-prone parts of speech than a noun. In terms of *Hypothesis 1*, what this proves is that diction does not change drastically in sentences where the subject is male or female when it comes to appearance as based on the coefficients these words are used interchangeably at a high level. Therefore, the trend of appearance words can be said to accept said hypothesis. But, based on the coefficient range, realistically-speaking, appearance words only make up a minor part of the bias exhibited in this linguistic analysis.

A closer look at the coefficients for the individual appearance words reveals that very few words in this category are exclusively male or female. This is in line with the rejection of *Hypothesis 1*. However, deviations from this overall trend do occur, and they are telling in that they reflect

gender stereotypes. For instance, the word ”beautiful” is exclusively used to describe women, and ”handsome” is only used when the subject of the sentence is male. The implication of this is, from a diction perspective, that word choice does change when describing a male or female’s looks. Interestingly, however, the coefficients for these descriptors seem to be evolving in an inverse fashion, with the coefficient for ”beautiful” becoming lower each year, and that of ”handsome” becoming higher. This seems to indicate that language is indeed changing over time.

Another insight, is that the word ”strong” is exclusively used in sentences where the subject is female. From a diction perspective, what this could imply is that there is a need to specify that women are strong while it is potentially an implied quality for men. This is in line with gendered notions of males and females. However, when juxtaposing ”strong” with ”muscular” in Figure 27, words used interchangeably in natural text, the overwhelmingly feminine trend of ”strong” is more sensical. As in, the text could genuinely be highlighting a women’s strength rather than using this term because it is implied stereotypically that women are not strong. This is supported by the fact that ”muscular” in Figure 27 started off as overwhelmingly masculine, but over time, it has been trending towards 0, indicating it has become less masculine.

#### 14.2.2.2 *Agency*

The notion of agency has been a core part of the feminist struggle. The first two waves of feminism were focused on raising awareness around the fact that women are as deserving of agency as men,

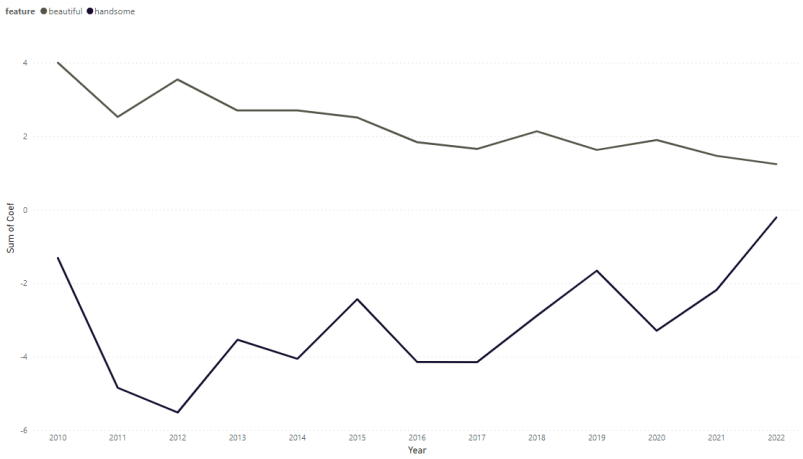


Figure 26: Trend of ”Beautiful” v. ”Handsome”

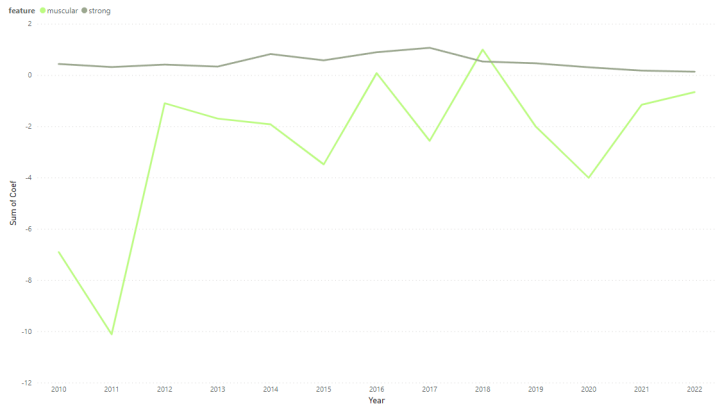


Figure 27: Trend of ”Strong” v. ”Muscular”

and that this was essential for their equal participation in socioeconomic life. Using (Lawson et al. 2022)’s agency word list found in Appendix 18.1.4, this study measures the degree to which agency words are associated to different genders.

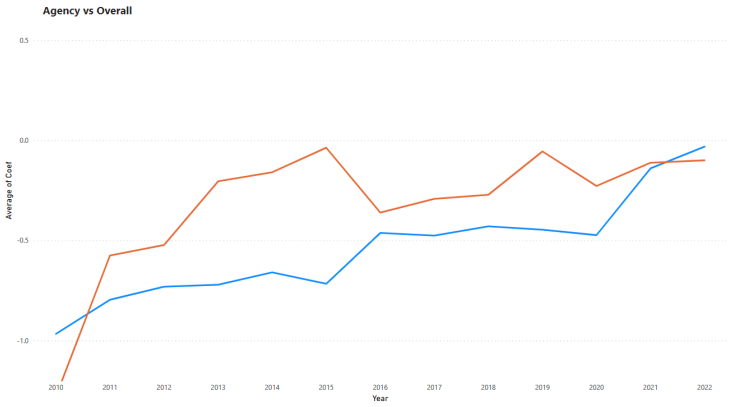


Figure 28: Trend of Agency Words vs. Overall Trend

Figure 28 shows that agency words are always masculine over the period studied. Overall, it can be said they exhibit a slight upward trend, however, it never reaches a coefficient of 0 or above. This entails that agency is mostly associated with the majority class - men. This is expected as per the literature reviewed in 6.1.5 and seems to confirm *Hypothesis 1* - that there is a change in word choice when speaking about men and women, particularly when it comes to agency.

The overall agency trend is explainable when auditing the coefficients for the individual words in this list over the years. As such, *Hypothesis 3* - that bias is exhibited and explainable - may not be rejected. Neither masculine or feminine related words have large magnitude coefficients. But over time, masculine agency words exhibit larger coefficients in terms of absolute value than

words associated to women. Examples of masculine agency words throughout 2010-2022 include ”leader”, ”command”, and ”power”. The contrast between the trends of ”power” and ”leader” shown in Figure 29 is interesting. While these words a priori share a proportional relationship, in that leaders tend to have more power, and power tends to be related to leadership, they take different directions over time. ”Power” has an upward trend, meaning over time it becomes less masculine. On the other hand, ”leader” is consistently low in the earlier part of the decade, has a significant rise from 2015-2020 and then dips back to 2010 levels by 2022. This relationship among agentic words shows that while some progress has been made, women are still not portrayed nowhere near as agentic as men in the media as hypothesized in *Hypothesis 1* and *Hypothesis 3*.


Figure 29: ”Leader” v. ”Power”

The only agency word which is consistently feminine is ”strong”. Other agency words do end up leaning female by 2022, but their magnitude is relatively small. These dynamics explain the overall dominion of masculine agency words in the study.

#### 14.2.2.3 *Communality*

An important part of the feminist discourse on gender bias discussed in 6.1.5 highlights that institutions such as the media play a role in perpetuating the misrepresentation of women. This serves to perpetuate the cultural imperialism that maintains the patriarchy as the dominant class in society. The misrepresentation of women has multiple facets, including behavioral stereotypes, which usually portray women as being communally-driven. The hypothesis based on this, is that

communality words should have a positive coefficient, thereby being associated as being more feminine. The communality words used in this analysis are from Lawson et al. (2022) and can be found in Appendix 18.1.2.

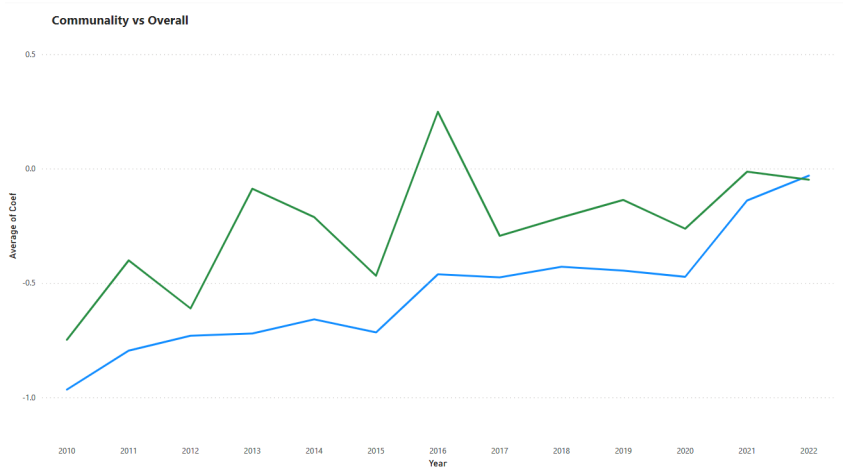


Figure 30: Trend of Communality Words vs. Overall Trend

Based on how communality trends above the overall coefficient benchmark, it can be said that communality words are more feminine than the average of other word list coefficients. However, as Figure 30 demonstrates, there is a ground to reject *Hypothesis 1* based on the coefficients of the overall communality trend. These tend to be negative - ergo, masculine - in all years studied but 2016. Given the nuanced nature of communality words, the bias is not explainable and therefore *Hypothesis 3* cannot be accepted.

A potential explanation for the masculine trend observed in communality words is that the nature of being communal can be political or social. The news, being a reflection of every day occurrences, is likely to capture both. The trend leans towards negative coefficients because it is an average of all coefficients of communality words. Upon closer inspection of individual words, it is notable that words that are considered more masculine have larger coefficients in terms of absolute value. Examples of these are ”cooperation”, ”polite”, ”truthful”, and ”moral”. Words feminist literature would deem a stereotypical misrepresentation of women such as ”social”, ”equality”, and ”love” all have positive coefficients, but these are smaller in absolute values than the masculine-leaning words, thereby not affecting the general trend as much. As such, while *Hypothesis 1* cannot be accepted at

a high-level, it does hold when comparing individual words from the viewpoint of stereotypical communality attributes.

#### 14.2.2.4 *Power*

Power words are defined in the Appendix 18.1.3 and in the Data Descriptions 8 where the methodology for collection and reasoning is highlighted. In this dataset power is seen as words that bring power to the agent identified in the sentence. This is reflected in words like repulse and boasts. When examining the trend line for power words in figure 31 the line is still never crosses the threshold to being a predictor for the female class. This overall trend and the coefficients present show that while the words are more masculine they are not great predictors of the masculine class overall. Despite the correlation being less strong it still supports *Hypothesis 3* as the overall category of Power words do help to predict the male class.

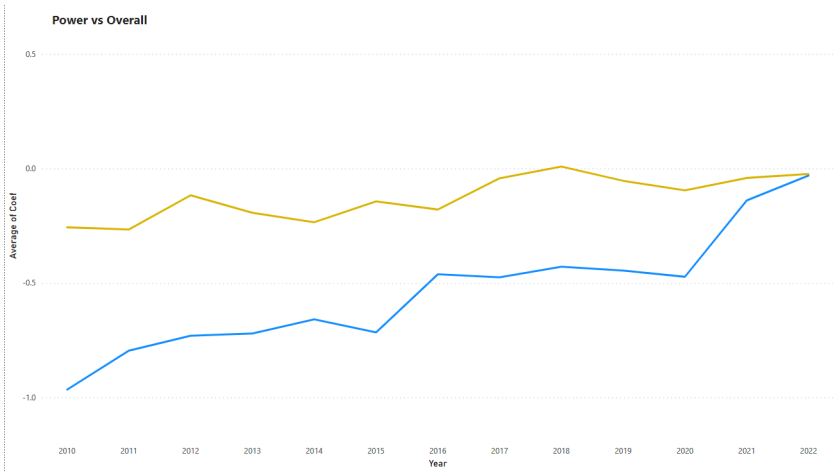


Figure 31: Overall trend of Power Words vs the Overall trend

Within the power classification there are some words that exclusively belong to the masculine and feminine class. These words are, similarly to the violence category, have different types of word in each classification. For women the words are more communal like ”listen” and ”trust” or products of victimization like ”stalk” and ”struggle”. For masculine words the power words seem to display more agency with words like ”admit” and ”offend.

#### 14.2.2.5 *Violence*

In order for the patriarchy to maintain the class imbalance between men and women, the system must perpetuate itself through any means necessary. This calls back to the five faces of oppression where violence is seen as a method to perpetuate the patriarchy. The Media can use violence as a tool for oppression by outright calling for violence, but also in much more subtle ways. While a woman is more likely to be a victim of a crime with a man as the perpetrator, The Media has the choice of what situations to cover, how frequently, and the adjectives that are used to describe the victim and perpetrator.

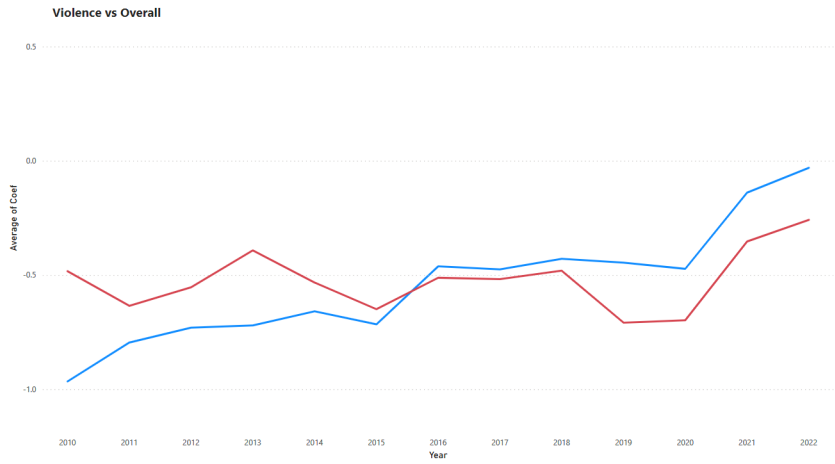


Figure 32: Overall trend of Violence Words vs the Overall trend

When examining the overall trend of the violence line, it's clear that the trend is significantly more masculine than feminine with the average coefficient of -.53 over the 12 year time period. This line is relatively stable with a standard deviation of 1.54. This overall line supports *Hypothesis 3* that this bias towards men is explainable and the conclusion that violent language is pervasive in the male class.

When the words are split into absolute class words (words that stay above or below 0 for the entire timeseries) there is more information about the violence that is associated exclusively with men and women. Almost all exclusively female words are associated with domestic violence, rape, and abuse. This compares to the male list where the words span categories of war (warhead, military, solider) perpetration (conviction, prison, crime), and calls to violence (cut, stab, injure).

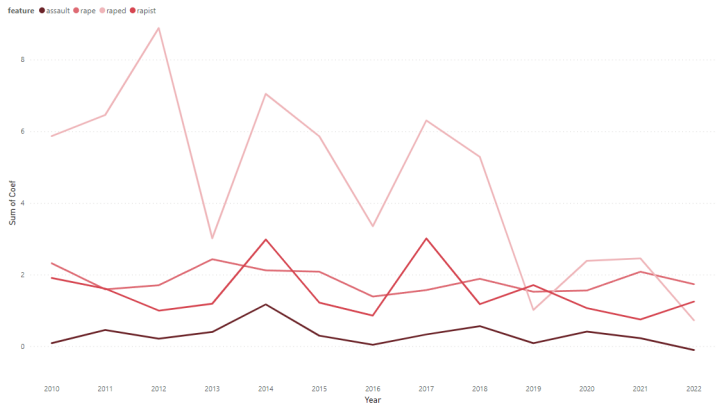


Figure 33: Rape, rapist, rapes, and assault trend line

| Male | Female |
|---|---|
| wound, weapon, war, violent, trouble, troop, trial, threat, tension, suspect, stab, soldier, shot, shoot, sentence, robbery, raid, radical, punch, prison, offence, military, knife, jihad, jail, injury, incident, imperialist, gun, gang, fight, explosive, enemy, dispute, die, defend, damage, cut, crime, convict, consequence, confront, conflict, charge, casualty, burglary, brawl, battle, bad, arrest, army, arm, altercation | abduct, affect, anxiety, bruise, complain, cry, devastate, distress, domestic, harass, hysterical, pain, person, rape, rapist, scream, sexual, tear, violence |

Table 6: Exclusively Male and Female Violence Terminology

The absolute values can be a seen as a sign of the perpetuation of rape culture. While rape and sexual assault are common occurrences for women globally, sexual assault takes two and the weight of rape, rapist, assault, and rape (See Figure 33 being exclusively female excludes the (most often) male perpetrator.

All of the rest of the violence words cross into both classes at least once during the timeseries. While in one year a violence word can be female, the trend of overall belonging to the male class says more about the dataset. This data can be influenced by crime trends, slang changes, and instances where the victim/perpetrator relationship flips.

### 14.2.3 *Occupational-pair Trends*

## 14.3 BI-LSTM RNN

Due to the experimental nature of the developed equation, word breakdown results are provided with the caveat of further testing needed to fully validate the results. Thus the subsequent section is less robust than the logistic regression. Also, the data has only been assessed on a by world level for the outlined lexicons in Section 8.

### 14.3.1 *Overall*

Compered to Logistic Regression, there is much less variability in the categories. This is somewhat expected as the results are normalized and each model produces independent results with it's own weights and considerations for the inputs. Despite the overall trendlines being less insightful, there are meaningful insights within the words trends and word trends in comparison.

The RNN seems to be more sensitive to the Pandemic, as every category's average weight drops significantly for 2020, with very few of the assessed words even being in the female class in 2020. This is not unsurprising considering that almost all "Celebrity" and "Style" sections decreased dramatically while "Science" and "Politics" (mostly masculine categories) soared. In 2020 the most feminine word, in the assessed dataset, was "trusting" and the lowest "affray".

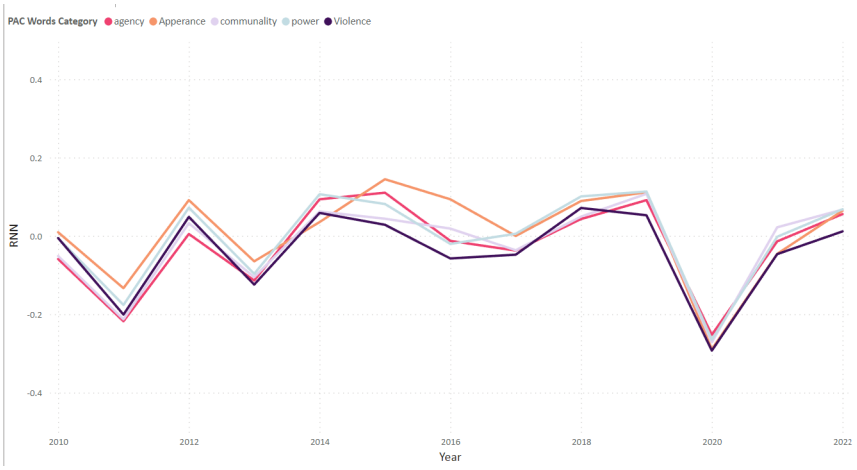


Figure 34: Average Normalized RNN Category Results

### 14.3.2 *Appearance*

Of all of the Lexicons, the Appearance words scored the highest (most feminine) for the RNN results. The words ”beautiful” and ”gorgeous” remaining positive for the entire timeseries. All other words crossed into the negative space at least once. One subtlety that the RNN picked up on was ”pretty” where it remained around zero for the entire time. This is, most likely, due to pretty being an adjective but also an adverb meaning ”a little bit”. The Logistic Regression model has scored the word as significantly more masculine than the RNN. Because the RNN takes word order into consideration, this leads to a greater ”intelligence” to the weights that are applied based on the learnings from the training set.

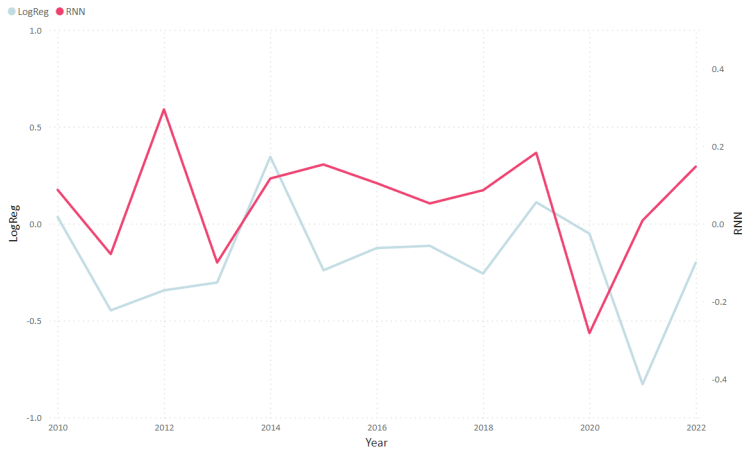


Figure 35: RNN Model vs Logistic Regression trend of ”Pretty”

### 14.3.3 *Agency*

Within the agency class, the adjective ”intelligent” and noun of ”intellectual” both exist. While conceptually these are close concepts, in reality they carry a deeper difference in meaning. In Figure 36 there is a strong trend in the gap between the two words with the lines flipping in 2017 and 2022.

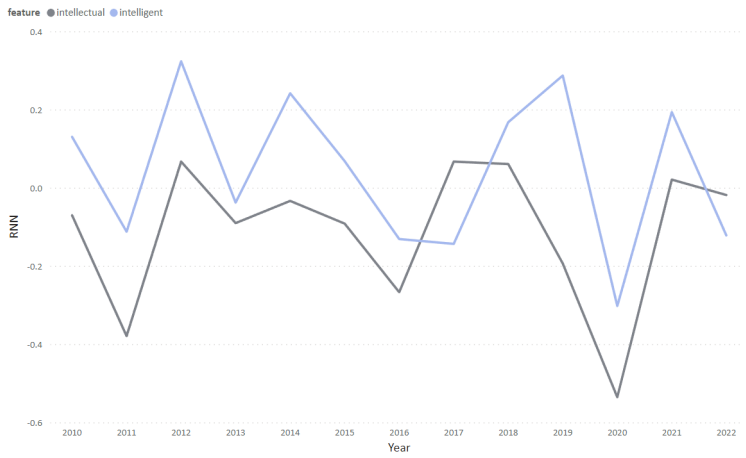


Figure 36: RNN Model Difference in ”Intelligent” and ”Intellectual”

### 14.3.4 *Communality*

Within the Bi-LSTM RNN results, the communality words skew more female, on average, than any other word category. The words tend to have large oscillations between years, very similar to logistic regression. The logistic regression and Bi-LSTM RNN have almost the same high and low ranking words on average with an exception of ”grieve” ranking much higher with the RNN for the female class and ”assist” having a much lower score. With the male class ”misunderstand” has a higher score than the logistic regression and so does ”repulse”. With the common turn of ”the grieving widow” and tropes about how ”men misunderstand everything” this development is surprising.

### 14.3.5 *Power & Violence*

All findings in the violence and power section for the RNN support the Logistic Regression. Some words changed their overall rank but these trends held between models, the violence words are overwhelmingly male and the power words are more in line with a center ranking, i.e. not a great indicator of gender as an entire category.

# 15

# DISCUSSION OF RESULTS

## 15.1 MACHINE LEARNING

Both the LR and Bi-LSTM RNN performed well in terms of macro accuracy and F1-score. This is significant given that this study looks at real-world data, and had a very limited preprocessing scope. This shows that they are capable of predicting class labels and the label scores show a balance between precision and recall. Therefore, the results analyzed in Section 14.3.5 are valid and representative. Performance could have certainly been higher had more preprocessing been carried out. However, a bare essentials approach was taken to prevent self-inflicted selection bias arising by removing too much data. The differences between the models are not attributable to preprocessing, though, as the same process was carried out for both.

The captured coefficients of the LR model give insights into the dataset that was used to train the model. The insights are seen in prior sections 13.2.2 comparing word scores. Accuracy and F1-score are affected by the fact that some features have low discriminatory power but have high probabilistic power individually. For example, ”assault” may be referred to highly for both classes, with sentences like ”He was convicted of sexual assault” and ”She was assaulted last July.” These words could have similar probabilities for both classes because of their nature with an aggressor and victim. Therefore the terms would have low discriminatory power. The discriminatory power issue could have been avoided by looking at verb tenses more closely or defining a threshold through which to eliminate features with low discriminatory power. The Bi-LSTM RNN is less sensitive to

this low discriminatory problem because it utilizes the context of the sentences to provide insights and outputs an independent score for each class. But, the Bi-LSTM RNN is still a deep learning ”black box” methodology so insights on word and terminology performance are largely hidden. This is why the *Difference Score* method was developed in Section 12.

Another important point with regards to the Bi-LSTM RNN model is that it seems to be sensitive to the volume of data, particularly when it comes to the female class. From 2020 onwards, the Bi-LSTM RNN macro accuracy and F1-scores drop substantially, almost reaching LR model levels. This volumetric change can be attributed to scraping methods but could also be due to Pandemic changes in employment and news coverage. While this is certainly indicative that change is occurring, it might also highlight that more data capturing the increased variance is necessary. Based on the micro precision and F1-scores for the female class, it can be said that this is more of an issue for the female class.

When analyzing the models in the Performance section 13.2.2 the macro-aggregate models were considered. But to fully evaluate performance the macro-aggregate methods must be contextualized with a micro-aggregate ones to comprehensively address the hypotheses in this study. This is showcased through the following example. The first brought forth the surprising result that the LR model had a higher macro accuracy than the Bi-LSTM RNN model in 2022. When looking at the micro figures, it is clear that the LR model exhibits a higher macro accuracy in 2022 because there is a higher precision score for the male class compared to the Bi-LSTM RNN - 80% and 70%, respectively. This highlights an important drawback of using macro methods: they can be influenced by higher magnitude scores to indicate a result that is not truly reflective of what is happening at the class-level. In the case of 2022, the male and female class exhibited a lower difference in precision scores in the Bi-LSTM RNN, while a larger one in the LR model, and this is thus what drove the latter to seemingly be more accurate at the macro level.

Triangulation among macro and micro aggregate methods must also be considered for the F1-score. The macro F1-scores had been trending downwards consistently since 2013, indicating that over time, there had been a change in the syntax in the text, the micro F1-scores for the different

classes showcase a different panorama. In fact, as seen in Figure 15, precisely in the years where the macro F1 scores start drifting downwards, some of the biggest differences in the micro F1 scores between the male and female classes are exhibited. This makes sense from a mathematical perspective, as it is explainable based on the macro F1-score equation, however insofar as the research question, it shows that a downward macro trend is not enough to accept or reject said hypothesis, in this case *Hypothesis 3*. This is because no improvement can be seen in the micro F1-score for the female class in addition to the overall trend. As such, macro metrics, like the micro ones, should be taken with a grain of salt when it comes to interpreting performance. Thus, an approach combining both methods is preferred.

## 15.2 IMPLICATIONS

#### 15.2.0.1 *The Media*

The news, as an institution, is meant to democratize true unbiased information to the masses. The Media provides this information through public channels to provide up-to-date information for readers. This information is intended to be factual and representative of the information. When examining the results from the study, it can be observed that although bias is declining, it does exist. This intrinsic bias is directly opposed to the objective of factual information. While some bias will always exist - the difference in women's reproductive health or rates of sexual assault - if this bias was intrinsic and not systematic the bias would remain level for words where women and men have equal stakes like "family" and "child". This would not remain completely female. The inability to reject *Hypothesis 1* and *Hypothesis 2* shows that fundamentally that one speaks differently in sentences involving men versus women. Ultimately these linguistic differences show that the BBC 50:50 project[1] is not doing enough.

1 See the problem formulation for a summary of the project

#### 15.2.0.2 *The System*

As stated in the Conceptual Framework 5, large news organizations are systems that operate independently of the people that make them. The BBC is part of the system of The Media where the organization must perpetuate itself and the system that it exists within. This means that without systemic change from outside and within, The Media will continue to oppress marginalized classes and participate in the perpetuation of the patriarchy. The BBC has tried to engage in systemic change with the 50:50 project where they reengaged with the concept of representation in their news room. While the BBC has shown that their representation in their news stories has reached equality, they have not achieved a state where bias has been excluded as proven in this study. Equality cannot be achieved when there is a difference in the descriptors and syntactic structures used when speaking about men and women. While it is the expectation that individuals, as participants in the system, will perpetuate the system from within, The Media is in control of the narrative that they produce. By engaging in biased language The Media is codifying gender bias through their language choices.

”Hysterical” stems from the diagnosis of Hysteria, an ailment now largely considered pseudo-science, in its allegations that women were ”too emotional”, ”too sexual”, or ”not sexual enough” and historically led to the institutionalization of women and unwarranted hysterectomies(Tasca et al. 2012). Hysterical is a call back term to the oppression of women utilizing the institution of medicine to harm women and yet it is still used in the news and is highly associated with the female class (See Figure 19). This is but one example of the descriptors associated with the oppression of women used by the BBC in their news articles in the years 2010-22. By utilizing terminology of this kind, The Media is an actively engaged in the oppression.

#### 15.2.0.3 *Business Implications*

Media agencies need to appeal to their audience, as this is how they fundamentally generate revenue and survive. The Media will always exist as a system, but the participating institutions within it can change. Bias identification, thus, has important business implications for news companies.

The patriarchy considers the Media an institutional oppressor. However, systemic change on issues such as equality are happening. Traditional institutions such as the BBC can either rely on systemic change to be changed or they can institutionally transform themselves in anticipation of the demands of a more equal system. The stakes are high because they way in which news agencies handle these systemic changes has not only ethical but also financial impacts. News sources' legitimacy, trust, and reputation are on the line, and these indicators are directly correlated with demand for their content. As such, working towards the end of rectifying gender bias in article content is key when devising future-proof business plans. This is because the consumers will newer generations - individuals who have largely participated in driving this systemic change - who are known to be more progressive, particularly on topics like representation.

It is a known fact that the BBC wants to increase representation. This is evinced by their "50:50 The Equality Project". While the 50:50 project is an important step forward in terms of increased representation, it is not enough to drive the kind of systemic change necessary to neutralize the misrepresentation of the non-dominant classes. To accomplish this, a change in the narrative is needed. An alternative the BBC should consider to this end, is to leverage machine learning - specifically NLP techniques - in the same vein as this paper, to understand the bias they perpetuate as a sociocultural institution. Doing so would enable them to rectify their future narratives to achieve a more comprehensive representation of women. This thesis proposes a foundational approach for this, using simple and complex patterns of bias at the word- and sentence-level using LR and Bi-LSTM RNN algorithms.

## 15.3 STUDY CAVEATS

While the study is robust in terms of number of documents used for training there are caveats to consider when assessing and understanding the results. While many statements have been made about the media holistically and the BBC is a central part of the media as a whole, the BBC is not The Media. This is important especially in light of the efforts that the BBC has made to be more representative of women in their reporting. Not all media organizations are doing so and as such

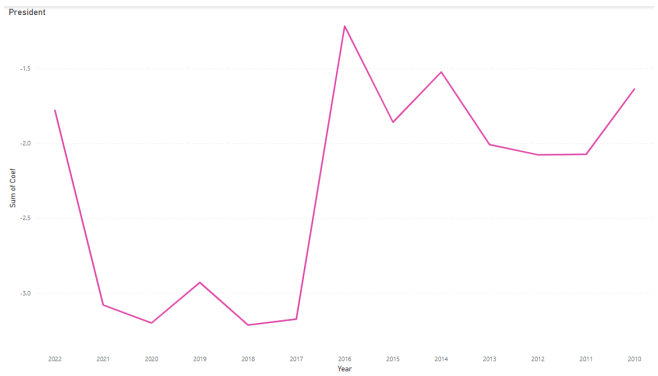


Figure 37: Trendline of ”president” Logistic Regression Coefficients

display different kinds of bias. This was explored in Asr et al. 2021 where Canadian media outlets were assessed on a contributor level to look for bias. This may seem counter-intuitive to points made in the paper but it is important to examine the BBC's role in The Media and how it influences The Media as a whole. Much like the dog that wags the tail, the BBC, the tail, may not wag the dog, The Media, but it is a part of the dog and it moves with the system as a whole or it is not part of the system.

News data also lives in the context of the news. In Figure 38 a small timeline has been overlayed on the timeline of the overall linear regression coefficients. This timeline explores some key events in feminist history and in UK history. They are key moments in history that systemic change has been written into history, which would undoubtedly cause a change in the narrative. This change in narrative is reflected in the data where ”president”, see Figure 37 reaches it's most feminine point in 2016 when Hilary Clinton ran for president.

Every year except for 2022 in the Logistic Regression passed *Hypothesis 1*. However, to definitely answer whether bias has decreased, a trend must exist, rather than just a point in time. Given that further into the future is a counterfactual, only time can tell if the changes of 2022 are the start of something new.

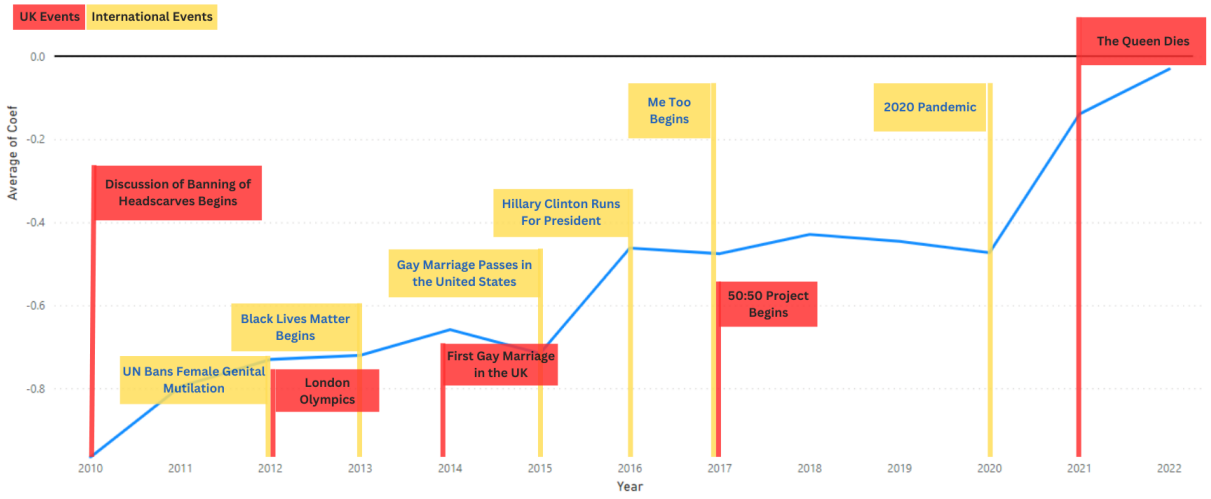


Figure 38: Timeline of Events between 2010-2022

# 16

# LIMITATIONS

The limitations are categorized according to the different parts of the methodological approach that they pertain to: data, preprocessing, algorithms, and result analysis.

## 16.1 DATA

There are two constraints regarding the data utilized in this study. Firstly, while voluminous, the data attained was not completely balanced between years and genders. As addressed in the Section 9, given that the underlying distribution between the male and female is relevant for the hypotheses we seek to prove, no class balancing algorithm was deemed necessary. Nonetheless, it should be acknowledged that said disproportion in class data might have impacted the algorithms' ability to predict results. This, is particularly evident given the conclusions reached about the Bi-LSTM RNN performance 13.2.2. Secondly, given that this thesis relied on a third-party web-scraping site to gather the data, it is not known for certain whether the data analyzed is all of the BBC news data possibly available for 2010 to 2022. The uncertainty of how representative the sample of data studied is, is a limitation in itself, regardless of how reliable the provider may be.

## 16.2 PRE-PROCESSING

The issue of leakage is a key limitation in terms of preprocessing. This is relevant because this study relies on a methodological approach where all gender identifiers - pronouns, first names, honorific titles and select noun-pairs - are removed from the data used to train and test the model.

The reason for this being that by removing words explicitly descriptive of gender relevant features related to the respective genders could be better identified. The leakage manifested itself in two ways: in first-names and select gendered nouns. To accomplish first-name removal the NLTK's named entity recognition library was utilized to identify the presence of a named entity. While most names were captured some first names like ”Malala” trickled through to the results. Then there is also the issue of first names which are also nouns or verbs, which this library would not have captured such as ”hope” and ”dawn”. These were only visible when examining all results in aggregate. An improvement that could be made in this regard, which has been put into practice in existing NLP scholarship (Asr et al. 2021), is to capture the top 1000 celebrities' first names and genders, and manually removed these from the data.

## 16.3 ALGORITHMS

By nature of their functioning, the algorithms utilized have limitations that should be considered. While the LR models are useful to analyze feature importance over time, through its coefficients, said algorithm is simple and unable to capture some of the more complex relationships between feature and target variables. This is reflected in the LR model performance, where its predictive power is shown to be lower than the Bi-LSTM RNN. Therefore, despite being useful to inform *Hypothesis 1* it does not contribute much in assessing *Hypothesis 3*.

The performance of the Bi-LSTM RNN models was stronger than that of the LR ones, proving they are more suitable to meet the demands of the research question, when it comes to *Hypothesis 3*. A notable constraint for this algorithm, however, is the skepticism that in business contexts towards ”black box” components in machine learning applications. Thus, while not a limitation in terms of answering the research question, it should be highlighted as one that might impact the business application of this paper's methodology.

## 16.4 APPROACH TO THE ANALYSIS OF RESULTS

A key thing to note about the results of this study is that the BBC is a news source, but not all the news. Therefore, no conclusions about The Media as a system can be drawn. The insights presented in this study are thus of a limited, yet defined, scope. Another notable element of this study's results is that the methodology used to speak to investigate *Hypothesis 2* is originally developed by the authors of this thesis. This method generated sensical test results and the algorithms deployed to this end had a strong performance. However, it has not been battle-tested in other the literature, as there is a overall gap in NLP scholarship on feature extraction for RNNs. Amidst the limited literature on the subject, studies deploying a similar methodology are outlined in 12.

Part VI

# CONCLUSION

# 17

# CONCLUSIVE REMARKS

This thesis investigated whether gender bias in language used in The Media has changed in the era of datafication, by using Logistic Regression and Bi-LSTM RNN algorithms. The study found that while there were positive changes observed across the 2010-2022 period studied, diction and syntactic choices still vary based on the gender of the subject in a sentence. The Bi-LSTM RNN models were particularly effective in identifying this bias.

The research has significant social and business implications, especially as younger generations who prioritize equality become more involved in public and private institutions. The media has a responsibility to accurately portray facts and combat systemic biases, yet it has historically contributed to the perpetuation of glass ceilings that limit women's opportunities. Language is a powerful tool for transmitting bias in news institutions, so understanding the patterns in text is crucial to identifying and addressing these issues.

Although the system of bias may persist, there is evidence of progressive, systemic change occurring. While this change is occurring, it is not enough. The bias that is still present is harmful and new generations will continue to demand change. News institutions must evolve to meet the demands of this metamorphosis, as their ability to do so will determine their survival. In the age of datafication, where data intensifies competition and accelerates institutional transformation, leveraging NLP to drive positive change through data offers a significant opportunity for impact beyond business, that can affect society and culture as a whole. This approach may be key to advancing ideals of equality in the context of AI and the implicit biases that still impede progress

for women today. The proliferation of AI and machine learning could easily lead to the next wave of feminism structured around making AI less biased and using it to detect and change implicit biases. The whole world has an opportunity to use new technology to reduce bias instead of perpetuating it.

# 18

# FUTURE WORK

Many elements of this thesis have been exploratory in nature. By building an accurate model for gender prediction, the findings from these coefficients and difference scores can built atop of to help capture implicit bias and change the linguistic paradigm. The simplest way to combat implicit bias is through the knowledge of it's existence.

Specified lexicons were highlighted in this paper as potential areas that contain implicit and explicit bias. Through the results within these lexicons, language choices that are strongly correlated to a specific class can be offered as opportunities to change word choices. Put simply, men are also beautiful and women are also leaders. These highly biased linguistic choices could be made into simplistic suggestions for writers to try and reduce their bias.

While simply substituting a word for another can just create the problem of new terminology being developed to circumvent the "rules" (see TikTok banning "suicide" so the word "unalive" was developed), creating a framework that picks up on contributing factors to implicit bias and detects them would solve this. The current results of this paper could be used to develop a deep learning method that tests and finds words with high difference scores and offers them as contributing factors to implicit bias. While this would require further testing of the Difference Score technique, it would potentially allow for a more sophisticated methodology to be used to capture bias, without being impacted by that bias.

To further study implicit bias, the models could be retrained on a dataset with all roles, biological nouns and sexual health markers removed, to observe what kind of adjectives and nouns remain as

high coefficients in the model. This would provide more insight into the linguistic choices that are truly neutral by nature.

The impact of the high coefficient descriptor words and actions could be removed in a secondary study where the presence or absence of these words could be observed in the f1 and accuracy measures to see the impact on the models ability to predict the class. This would provide further proof in the level of bias that the word portrays.

To clarify results from this paper, many other corpera could be studied with the same techniques and the subsequent results could then be compared. This could be done within the news but also in other fields like blog posts, research papers, or any other publication with long form text. While a similar process could be done on short from text, we feel that implicit bias is much more pervasive and important in long form. As generative AI continues to dominate the landscape, the discussion of bias will only increase as the prevalence of AI will cause the perpetuation of these biases.

# BIBLIOGRAPHY


Acker, Joan (1992). “Gendering organizational theory”. In: *Classics of organizational theory* 6, pp. 450–459.

Adams, Maurianne and Lee Anne Bell (May 2007). *Teaching for Diversity and Social Justice*. Ed. by Pat Griffin. 2nd ed. London, England: Routledge.

Aggarwal, Komal et al. (2022). “Deep learning in robotics for strengthening industry 4.0: opportunities, challenges and future directions”. In: *Robotics and AI for Cybersecurity and Critical Infrastructure in Smart Cities*, pp. 1–19.

Ahmed, Mostaq, Partha Chakraborty, and Tanupriya Choudhury (2022). “Bangla document categorization using deep RNN model with attention mechanism”. In: *Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021*. Springer, pp. 137–147.

Ahnve, Fredrik et al. (2020). “Predicting Stock Price Movements with Text Data using Labeling based on Financial Theory”. English. In: *Proceedings - 2020 IEEE International Conference on Big Data. Big Data 2020*. Ed. by Xintao Wu et al. null ; Conference date: 10-12-2020 Through 13-12-2020. United States: IEEE, pp. 4365–4372. ISBN: 9781728162522. DOI: 10.1109/BigData50022.2020.9378054. URL: https://bigdataieee.org/BigData2020/.

Akhtar, Md. Shad et al. (2017). “Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis”. In: *Knowledge-Based Systems* 125, pp. 116–135.

Alsharef, Ahmad et al. (2022). “An automated toxicity classification on social media using LSTM and word embedding”. In: *Computational Intelligence and Neuroscience* 2022.

Alvarez-Melis, David and Martin Saveski (2016). “Topic modeling in Twitter: Aggregating tweets by conversations”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10. 1, pp. 519–522.

Asr, Fatemeh Torabi et al. (2021). “The gender gap tracker: Using natural language processing to measure gender bias in media”. In: *PloS one* 16.1, e0245533.

Bailey, Alexis H., Alex Williams, and Andrei Cimpian (2022). “Based on billions of words on the internet, people= men”. In: *Science Advances* 8.13, eabm2463.

Balter, Michael (Jan. 2015). “Human language may have evolved to help our ancestors make tools”. In: *Science*.

Bangyal, Waqas Haider et al. (2021). “Detection of fake news text classification on COVID-19 using deep learning approaches”. In: *Computational and Mathematical Methods in Medicine* 2021, pp. 1–14.

Bartl, Marion, Malvina Nissim, and Albert Gatt (2020). “Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias”. In: *CoRR* abs/2010.14534.

Bigler, Rebecca S. and Campbell Leaper (2015). “Gendered language: Psychological principles, evolving practices, and inclusive policies”. In: *Policy Insights from the Behavioral and Brain Sciences* 2.1, pp. 187–194.

Birjali, Mohammed, Mohammed Kasri, and Abdelghani Beni-Hssane (2021). “A comprehensive survey on sentiment analysis: Approaches, challenges and trends”. In: *Knowledge-Based Systems* 226, p. 107134.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet allocation”. In: *Journal of Machine Learning Research* 3.01, pp. 993–1022.

Bo, Pang and Lillian Lee (2008). *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval.

Bolukbasi, Tolga et al. (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems*. Vol. 29.

Boroditsky, Lera (2011). “How language shapes thought”. In: *Scientific American* 304.2, pp. 62–65.

Butler, Judith (1988). “Performative acts and gender constitution: An essay in phenomenology and feminist theory”. In: *Theatre Journal* 40.4, pp. 519–531.

Butler, Judith and G. Trouble (1990). “Feminism and the Subversion of Identity”. In: *Gender Trouble* 3.1.

Cai, Liping and Yixin Zhu (2015). “The challenges of data quality and data quality assessment in the big data era”. In: *Data Science Journal* 14.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017). “Semantics derived automatically from language corpora contain human-like biases”. In: *Science* 356.6334, pp. 183–186.

Chaloner, Kaytlin and Alfredo Maldonado (Aug. 2019). “Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, pp. 25–32. DOI: 10.18653/v1/W19-3804. URL: https://aclanthology.org/W19-3804.

Chao, Zehan et al. (2022). *Inference of Media Bias and Content Quality Using Natural-Language Processing*. arXiv: 2212.00237 [physics.soc-ph].

Chater, Nick and Morten H. Christiansen (2010). “Language acquisition meets language evolution”. In: *Cognitive Science* 34.7, pp. 1131–1157.

Chen, Wenhu et al. (2019). “How large a vocabulary does text classification need? a variational approach to vocabulary selection”. In: *arXiv preprint arXiv:1902.10339*.

Chen, Yunhao and Fei Pan (2022). “Multimodal Detection of Hateful Messages Using Visual-Linguistic Pre-Trained Deep Learning Models”. In: *arXiv preprint arXiv:2202.05336*.

Cheng, Liang et al. (2021). “Modeling temporal patterns of cyberbullying detection with hierarchical attention networks”. In: *ACM/IMS Transactions on Data Science* 2.2, pp. 1–23.

Chiranjeevi, P., D. Teja Santosh, and B. Vishnuvardhan (2019). “Survey on sentiment analysis methods for reputation evaluation”. In: *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*. Springer Singapore, pp. 53–66.

Chiu, Jason P. C. and Eric Nichols (2016). "Named entity recognition with bidirectional LSTM-CNNs". In: *Transactions of the Association for Computational Linguistics* 4, pp. 357–370.

Chomsky, Noam (2014). *Aspects of the Theory of Syntax*. Vol. 11. MIT Press.

Chomsky, Noam (2002). *Syntactic structures*. Mouton de Gruyter.

Christ, Carol P. (2016). "A new definition of patriarchy: Control of women's sexuality, private property, and war". In: *Feminist Theology* 24.3, pp. 214–225.

Crenshaw, Kimberlé (1989). "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics". In: *U. Chi. Legal F.* 139.

Davidson, Thomas et al. (2017). "Automated hate speech detection and the problem of offensive language". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1, pp. 512–515.

De Beauvoir, Simone (1949). "Woman as other". In: *1999), Social Theory*, pp. 337–339.

De Beauvoir, Simone and H. M. Moinaux (1953). *The Second Sex*.

Dessi, Danilo et al. (2021). "TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study". In: *arXiv preprint arXiv:2105.09632*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dong, Shihao, Peng Wang, and Khizar Abbas (2021). "A survey on deep learning and its applications". In: *Computer Science Review* 40, p. 100379.

Drikvandi, Roozbeh and Olufemi Lawal (2023). "Sparse principal component analysis for natural language processing". In: *Annals of Data Science* 10.1, pp. 25–41.

Drus, Zakaria and Haslina Khalid (2019). "Sentiment analysis in social media and its application: Systematic literature review". In: *Procedia Computer Science* 161, pp. 707–714.

Facio, Alda (2018). "The Parable of the Origin of Patriarchy author". In: *Canadian Woman Studies* 33.1–2.

Fang, Xing and Justin Zhan (2015). “Sentiment analysis using product review data”. In: *Journal of Big Data* 2.1, pp. 1–14.

Fernandez-Martinez, Juan L and Zulima Fernandez-Muniz (2020). “The curse of dimensionality in inverse problems”. In: *Journal of Computational and Applied Mathematics* 369, p. 112571.

Formanowicz, Magdalena and Karolina Hansen (2022). “Subtle linguistic cues affecting gender in (equality)”. In: *Journal of Language and Social Psychology* 41.2, pp. 127–147.

Fraser, Nancy (1990). “Rethinking the public sphere: A contribution to the critique of actually existing democracy”. In: *Social Text* 25/26, pp. 56–80.

French, Lisa (2014). “Gender then, gender now: Surveying women’s participation in Australian film and television industries”. In: *Continuum* 28.2, pp. 188–200.

Fricker, Miranda and Jennifer Hornsby, eds. (2000). *The Cambridge companion to feminism in philosophy*. Cambridge University Press.

Friedan, Betty (1963). *The feminine mystique*. WW Norton & Company.

Garg, Nikhil et al. (2018). “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Go, Alec, Lei Huang, and Richa Bhayani (2009). “Twitter sentiment analysis”. In: *Entropy* 17, p. 252.

Goffman, Erving (1981). *Forms of talk*. University of Pennsylvania Press.

Goldberg, Yoav (2017). *Neural network methods for natural language processing*. Vol. 10. 1, pp. 1–309.

Gonen, Hila and Yoav Goldberg (2019). “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them”. In: *arXiv preprint arXiv:1903.03862*.

Goodhew, Stephanie C et al. (2022). “The content of gender stereotypes embedded in language use”. In: *Journal of Language and Social Psychology* 41.2, pp. 219–231.

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks* 18.5-6, pp. 602–610.

Grus, Joel (2019). *Data Science from Scratch: First Principles with Python*. O'Reilly Media.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.

Hao, Xiaoke, Guangquan Zhang, and Shu Ma (2016). "Deep learning". In: *International Journal of Semantic Computing* 10.03, pp. 417–439.

Hassan, Syed Usama, Javed Ahamed, and Khaleeq Ahmad (2022). "Analytics of machine learning-based algorithms for text classification". In: *Sustainable Operations and Computers* 3, pp. 238–248.

Hauser, Marc D. et al. (2014). "The mystery of language evolution". In: *Frontiers in Psychology* 5, p. 401.

Heldke, Lisa Maree and Peg O'Connor (2004). *Oppression, privilege, and resistance: Theoretical perspectives on racism, sexism, and heterosexism*. McGraw-Hill Humanities, Social Sciences & World Languages.

Hovy, Dirk et al., eds. (Apr. 2017). *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics. DOI: 10.18653/v1/W17-16. URL: https://aclanthology.org/W17-1600.

Høysæter, Lars Smørås and Pål-Christian S Njølstad (2014). "Sentiment analysis for financial applications". In: *Norwegian University of Science and Technology*.

Itoo, Feroz and Satinder Singh (2021). "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection". In: *International Journal of Information Technology* 13, pp. 1503–1511.

Jin, Zhijiing (2021). URL: https://nlp4sg.vercel.app/.

Jing, Longfei and Yu Tian (2020). "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11, pp. 4037–4058.

Johnson, Allan G. (2004). “Patriarchy, the system”. In: *Women’s lives: Multicultural perspectives*. 3rd ed., pp. 25–32.

Jurafsky, Daniel and James Martin (2021). *Speech and Language Processing*. 2nd. Pearson Education.

Kagan, Danielle, Tyler Chesney, and Michael Fire (2020). “Using data science to understand the film industry’s gender gap”. In: *Palgrave Communications* 6.1, pp. 1–16.

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom (2014). “A convolutional neural network for modelling sentences”. In: *arXiv preprint arXiv:1404.2188*.

Kamath, Uday, Jia Liu, and James Whitaker (2019). *Deep Learning for NLP and Speech Recognition*. Vol. 84. Cham, Switzerland: Springer.

Kashid, Shamal et al. (2023). “Bi-RNN and Bi-LSTM Based Text Classification for Amazon Reviews”. In: *Key Digital Trends in Artificial Intelligence and Robotics: Proceedings of 4th International Conference on Deep Learning, Artificial Intelligence and Robotics,(ICDLAIR) 2022-Progress in Algorithms and Applications of Deep Learning*. Springer, pp. 62–72.

Kazar, Gizem et al. (2022). “Quality Failures–Based Critical Cost Impact Factors: Logistic Regression Analysis”. In: *Journal of Construction Engineering and Management* 148.12, p. 04022138.

Kimmel, Michael S., Amy Aronson, and Amy Kaler, eds. (2008). *The gendered society reader*. New York, NY: Oxford University Press.

Kiritchenko, Svetlana and Saif M. Mohammad (2018). “Examining gender and race bias in two hundred sentiment analysis systems”. In: *arXiv preprint arXiv:1805.04508*.

Kite, Mary E., Kay Deaux, and Elizabeth L. Haines (2008). “Gender stereotypes”. In: *Psychology of women: A handbook of issues and theories*. Ed. by Florence L. Denmark and Michele A. Paludi. Praeger Publishers/Greenwood Publishing Group, pp. 205–236.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019). “The geometry of culture: Analyzing the meanings of class through word embeddings”. In: *American Sociological Review* 84.5, pp. 905–949.

Lawson, Michael A. et al. (2022). “Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language”. In: *Proceedings of the National Academy of Sciences* 119.9, e2026443119.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *Nature* 521.7553, pp. 436–444.

Liang, Percy (2005). “Semi-supervised learning for natural language”. PhD thesis. Massachusetts Institute of Technology.

Liang, Shengbin et al. (2020). “A Double Channel CNN-LSTM Model for Text Classification”. In: *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, pp. 1316–1321.

Liu, Bing (2012). *Sentiment analysis and opinion mining*. Vol. 5. 1, pp. 1–167.

Lorber, Judith (2001). *Gender inequality*. Los Angeles, CA: Roxbury.

Lorber, Judith (1994). *Paradoxes of gender*. Yale University Press.

Lorber, Judith, Susan A Farrell, et al. (1991). *The social construction of gender*. Sage Newbury Park, CA.

Mabokela, Kebalepile R. and Thilo Schlippe (2022). “AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa”. In: *Artificial Intelligence Research: Third Southern African Conference, SACAIR 2022, Stellenbosch, South Africa, December 5–9, 2022, Proceedings*. Springer Nature Switzerland, pp. 309–322.

MacIver, Robert M. (1931). *Society: its structure and changes*. R. Long & RR Smith, Incorporated.

Markapudi, Baburao, Kunchaparthi Jyothsna Latha, and Kavitha Chaduvula (2021). “A New hybrid classification algorithm for predicting customer churn”. In: *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*. IEEE, pp. 1–4.

McConnell-Ginet, Sally (2014). “Meaning-Making and Ideologies of Gender and Sexuality”. In: *The handbook of language, gender, and sexuality*, pp. 316–334.

Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras (2006). “Spam filtering with naive Bayes-which naive Bayes?” In: *CEAS*. Vol. 17, pp. 28–69.

Mikolov, Tomas et al. (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Moghar, Adil and Mhamed Hamiche (2020). “Stock market prediction using LSTM recurrent neural network”. In: *Procedia Computer Science* 170, pp. 1168–1173.

Money, John, John L Hampson, and Joan G Hampson (Apr. 1960). “Hermaphroditism: Psychology & case management”. en. In: *Can. Psychiatr. Assoc. J.* 5.2, pp. 131–133.

Muller, Andreas C. and Sarah Guido (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Inc.

Murtagh, Fionn (2018). “Named Entity Recognition for Business Intelligence: Comparing Support Vector Machines and Recurrent Neural Networks”. In.

Nangia, Nikita et al. (2020). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. In: *CoRR* abs/2010.00133.

News, BBC (2023). *50:50 Project Methodology*. URL: https://www.bbc.co.uk/5050/methodology/ (visited on 03/05/2023).

News, BBC (2021). *BBC Reaches Record Global Audiences*. URL: https://www.bbc.co.uk/mediacentre/2021/bbc-reaches-record-global-audience (visited on 03/05/2023).

Nissim, Malvina, Rik van Noord, and Rob van der Goot (2019). “Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor”. In: *CoRR* abs/1905.09866. arXiv: 1905.09866. URL: http://arxiv.org/abs/1905.09866.

Olimov, Bekhzod et al. (2021). “Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model”. In: *Concurrency and Computation: Practice and Experience* 33.22, e6143.

Omvedt, Gail (1987). "The Origin of Patriarchy [Review of The Creation of Patriarchy, by G. Lerner]". In: *Economic and Political Weekly* 22.44, WS70–WS72. URL: http://www.jstor.org/stable/4377665.

Pankajakshan, R., S. Shafiq, and D. K. Dey (2021). "Emotion Detection in Financial Markets using News Headlines and TF-IDF". In.

Pranckevičius, Tadas and Vidmantas Marcinkevičius (2017). "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification". In: *Baltic Journal of Modern Computing* 5.2, p. 221.

Prewitt-Freilino, Jennifer L., Tiffani A. Caswell, and Emily K. Laakso (2012). "The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages". In: *Sex roles* 66.3, pp. 268–281.

Rácz, Anita, Dávid Bajusz, and Károly Héberger (2021). "Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification". In: *Molecules* 26.4, p. 1111.

Rahimi, Hamid (2019). "NLP-based Chatbots for Customer Service: A Review of Commercial Solutions". In.

Ramos, Juan (Dec. 2003). "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1, pp. 29–48.

Rasool, Ayesha et al. (2019). "Twitter sentiment analysis: A case study for apparel brands". In: *Journal of Physics: Conference Series*. Vol. 1176. 2. IOP Publishing, p. 022015.

Recasens, Marta, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky (2013). "Linguistic models for analyzing and detecting biased language". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1650–1659.

Rowe-Finkbeiner, Kristin (2004). *The F-Word: Women, politics, and the future*.

Rubin, Gayle (1975). "The traffic in women: Notes on the "political economy" of sex". In.

Sahagian, Graham (Dec. 2020). *What is Random State 42?* URL: https://grsahagian.medium.com/what-is-random-state-42-d803402ee76b#:~:text=

The%5C%20number%5C%2042%5C%20is%5C%20sort,over%5C%20the%5C%20period%5C%20of%5C%207.5.

Salton, Gerard (1975). "A vector space model for information retrieval". In: *Journal of the ASIS*, pp. 613–620.

Sap, Maarten et al. (2017). "Connotation frames of power and agency in modern films". In: *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2329–2334.

Sapir, Edward (1929). "The status of linguistics as a science". In: *Language*, pp. 207–214.

Schrupp, Antje (2017). *A brief history of feminism*. MIT Press.

Setiawan, Iwan et al. (2020). "HR analytics: Employee attrition analysis using logistic regression". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 830. 3. IOP Publishing, p. 032001.

Shah, Khyati et al. (2020). "A comparative analysis of logistic regression, random forest and KNN models for text classification". In: *Augmented Human Research* 5, pp. 1–16.

Sinha, Koustuv et al. (2021). "Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little". In: *CoRR* abs/2104.06644. arXiv: 2104.06644. URL: https://arxiv.org/abs/2104.06644.

Socher, Richard et al. (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Sparck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 28.1, pp. 11–21.

Tasca, Cecilia et al. (Oct. 2012). "Women And Hysteria In The History Of Mental Health". In: *Clinical Practice &amp Epidemiology in Mental Health* 8.1, pp. 110–119. DOI: 10.2174/1745017901208010110. URL: https://doi.org/10.2174/1745017901208010110.

Teele, Dawn L., Joshua Kalla, and Frances Rosenbluth (2018). "The ties that double bind: Social roles and women's underrepresentation in politics". In: *American Political Science Review* 112.3, pp. 525–541.

TensorFlow (2021). *Data Performance*. https://www.tensorflow.org/guide/data_performance. Accessed on May 1, 2023.

Tubishat, Mohammed, Norisma Idris, and Mohammad A. Abushariah (2018). "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges". In: *Information Processing & Management* 54.4, pp. 545–563.

Unger, Rhoda K. and Mary Crawford (1993). "Sex and gender—The troubled relationship between terms and concepts". In: *Psychological Science* 4.2, pp. 122–124.

Vaidhyanathan, Siva (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Van Huynh, Tin et al. (2019). "Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model". In: *arXiv preprint arXiv:1911.03644*.

Vangara, Raviteja et al. (2020). "Semantic nonnegative matrix factorization with automatic model determination for topic modeling". In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 328–335.

Vasilev, Ivan et al. (2019). *Python Deep Learning: Exploring Deep Learning Techniques and Neural Network Architectures with PyTorch, Keras, and TensorFlow*. Packt Publishing Ltd.

Vegt, Isabelle van der et al. (Mar. 2021). "The Grievance Dictionary: Understanding threatening language use". In: *Behavior Research Methods* 53.5, pp. 2105–2119. DOI: 10.3758/s13428-021-01536-2. URL: https://doi.org/10.3758/s13428-021-01536-2.

Wang, Peilin et al. (2020). "Classification of proactive personality: Text mining based on weibo text and short-answer questions text". In: *Ieee Access* 8, pp. 97370–97382.

Ward, L. Monique and Peggy Grower (2020). "Media and the development of gender role stereotypes". In: *Annual Review of Developmental Psychology* 2, pp. 177–199.

Weaver, William (1955). “A statistical approach to machine translation”. In: *Computational Linguistics* 16, pp. 79–85.

Wei, C., Y. Zhao, and X. Huang (2021). “A Systematic Comparison of Keyword Extraction Techniques from Scholarly Big Data”. In.

Wendland, Anna, Marco Zenere, and Jannis Niemann (2021). “Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique”. In: *Systems, Software and Services Process Improvement: 28th European Conference, EuroSPI 2021, Krems, Austria, September 1–3, 2021, Proceedings*. Vol. 28. Springer International Publishing, pp. 289–300.

Whorf, Benjamin Lee (1956). *Language, thought, and reality: Selected writings of...* Ed. by John B. Carroll.

Wolf, Naomi (2013). *The beauty myth: How images of beauty are used against women*. Random House.

Xu, Bing et al. (2015). *Empirical Evaluation of Rectified Activations in Convolutional Network*. arXiv: 1505.00853 [cs.LG].

Yang, Yiming and Xin Liu (Aug. 1999). “A re-examination of text categorization methods”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49.

Yao, Liang, Chengsheng Mao, and Yuan Luo (2019). “Graph Convolutional Networks for Text Classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 7370–7377.

Yeung, Ray M. and Wendy Yee (2011). “Logistic Regression: An advancement of predicting consumer purchase propensity”. In: *The Marketing Review* 11.1, pp. 71–81.

Young, Iris Marion (1990). “Throwing like a girl and other essays in feminist philosophy and social theory”. In.

Yu, Yong et al. (2019). “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures”. In: *Neural Computation* 31.7, pp. 1235–1270.

Yurynets, Roman et al. (2019). “Risk Assessment Technology of Crediting with the Use of Logistic Regression Model”. In: *COLINS*, pp. 153–162.

Zaidi, Syed Saqib Ali et al. (2022). “A survey of modern deep learning based object detection models”. In: *Digital Signal Processing*, p. 103514.

Zamudio, Margaret M and Francisco Rios (2006). “From traditional to liberal racism: Living racism in the everyday”. In: *Sociological Perspectives* 49.4, pp. 483–501.

Zhang, H., H. Li, and L. Huang (2021). “A Comparative Study of Information Retrieval Methods for Large-scale Text Data”. In.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). “Character-level convolutional networks for text classification”. In: *Advances in neural information processing systems* 28.

Zhang, Ze, Derek Robinson, and Jonathan Tepper (2018). “Detecting hate speech on twitter using a convolution-gru based deep neural network”. In: *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer International Publishing, pp. 745–760.

Zhao, Jieyu et al. (2019). “Gender Bias in Contextualized Word Embeddings”. In: *CoRR* abs/1904.03310. arXiv: 1904.03310. URL: http://arxiv.org/abs/1904.03310.

Zuheros, Carlos et al. (2021). “Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews”. In: *Information Fusion* 68, pp. 22–36.

## 18.1 CLASSIFICATION LEXICONS

### 18.1.1 *Appearance Words*

”alluring”, ”voluptuous”, ”blushing”, ”homely”, ”plump”, ”sensual”, ”gorgeous”, ”slim”, ”bald”, ”athletic”, ”fashionable”, ”stout”, ”ugly”, ”muscular”, ”slender”, ”feeble”, ”handsome”, ”healthy”, ”attractive”, ”fat”, ”weak”, ”thin”, ”pretty”, ”beautiful”, ”strong”

### 18.1.2 *Communality Words*

”accepting”, ”affectionate”, ”agreeable”, ”altruism”, ”altruistic”, ”attachment”, ”belonging”, ”benevolence”, ”caretake”, ”caring”, ”cheerful”, ”civility”, ”closeness”, ”communal”, ”communicative”, ”compassion”, ”compassionate”, ”compromising”, ”connected”, ”connections”, ”conscientious”, ”considerate”, ”consideration”, ”cooperation”, ”cooperative”, ”dependable”, ”dependency”, ”dependent”, ”duty”, ”easygoing”, ”enthusiastic”, ”equality”, ”fair”, ”faithful”, ”flexible”, ”forgiveness”, ”forgiving”, ”friendliness”, ”friendly”, ”generous”, ”gentle”, ”good”, ”gracious”, ”harmony”, ”helpful”, ”honest”, ”honesty”, ”hospitable”, ”humble”, ”humility”, ”influence”, ”interdependent”, ”interpersonal”, ”just”, ”kind”, ”love”, ”loyal”, ”loyalty”, ”moral”, ”nice”, ”nurture”, ”obliging”, ”optimistic”, ”patient”, ”pleasant”, ”polite”, ”politeness”, ”popular”, ”reasonable”, ”reliable”, ”respectful”, ”selfless”, ”sensitivity”, ”sincere”, ”sociable”, ”social”, ”supportive”, ”sympathetic”, ”talkative”, ”thoughtful”, ”tolerant”, ”trust”, ”trusting”, ”trustworthy”, ”truthful”, ”understanding”, ”universalism”, ”warm”, ”warmth”, ”welcoming”, ”wise”

### 18.1.3 *Power Words*

”worship”, ”apologize”, ”assist”, ”behold”, ”browse”, ”chase”, ”collect”, ”comprehend”, ”concentrate”, ”cram”, ”delight”, ”desire”, ”elaborate”, ”envy”, ”fill”, ”hail”, ”jot”, ”lug”, ”mime”, ”mimic”, ”nods”, ”nuzzle”, ”outline”, ”peruse”, ”please”, ”polish”, ”reach”, ”recall”, ”recognize”, ”reflect”, ”reiterate”, ”relieve”, ”relishe”, ”repulse”, ”require”, ”touch”, ”underestimate”, ”adorn”, ”aggravate”, ”amuse”, ”ascend”, ”brief”, ”dread”, ”dwell”, ”eyeballs”, ”fancy”, ”follow”, ”hate”, ”imitate”, ”miss”, ”misunderstand”, ”radiate”, ”repay”, ”respect”, ”salute”, ”sense”, ”silhouette”,

”stalk”, ”struggle”, ”subtitle”, ”accompany”, ”brush”, ”chant”, ”implore”, ”lack”, ”massage”, ”regard”, ”repeat”, ”seek”, ”shower”, ”smell”, ”tail”, ”unload”, ”weather”, ”believe”, ”compensate”, ”compile”, ”detect”, ”honor”, ”incorporate”, ”indulge”, ”like”, ”mingle”, ”adore”, ”applaud”, ”donate”, ”entertain”, ”evade”, ”grieve”, ”join”, ”mention”, ”oblige”, ”represent”, ”congratulate”, ”fulfill”, ”obey”, ”quote”, ”satisfy”, ”serve”, ”wonder”, ”credit”, ”defer”, ”echo”, ”emit”, ”seat”, ”cites”, ”disappoint”, ”experience”, ”value”, ”wish”, ”consult”, ”dream”, ”endure”, ”enforce”, ”enjoy”, ”learn”, ”offend”, ”pursue”, ”admire”, ”answer”, ”commemorates”, ”contract”, ”duck”, ”practice”, ”regret”, ”resent”, ”shadow”, ”undertake”, ”achieve”, ”cheer”, ”depart”, ”need”, ”pray”, ”promote”, ”visit”, ”beg”, ”copy”, ”plead”, ”tip”, ”uphold”, ”benefit”, ”celebrate”, ”constitute”, ”greet”, ”hope”, ”interest”, ”lose”, ”mind”, ”portray”, ”predict”, ”undergo”, ”vote”, ”want”, ”admit”, ”announce”, ”appeal”, ”ask”, ”attend”, ”await”, ”become”, ”call”, ”concede”, ”contribute”, ”explain”, ”fan”, ”fear”, ”guard”, ”hear”, ”listen”, ”note”, ”owe”, ”prefer”, ”receive”, ”recognize”, ”respond”, ”suffer”, ”thank”, ”trace”, ”trust”, ”witness”

### 18.1.4 *Agency Words*

”achievement”, ”active”, ”adamant”, ”aggressive”, ”ambition”, ”ambitious”, ”analytical”, ”assert”, ”assertive”, ”assertiveness”, ”assured”, ”autonomous”, ”autonomy”, ”bold”, ”bossy”, ”brave”, ”brilliant”, ”capability”, ”capable”, ”clever”, ”command”, ”competence”, ”competent”, ”competitive”, ”competitiveness”, ”confidence”, ”confident”, ”convincing”, ”creative”, ”cunning”, ”daring”, ”decisive”, ”determined”, ”diligent”, ”direct”, ”dominant”, ”dynamic”, ”educated”, ”effective”, ”efficient”, ”egocentric”, ”energetic”, ”exploration”, ”fast”, ”freedom”, ”hardworking”, ”imaginative”, ”independence”, ”independent”, ”individual”, ”individualistic”, ”industrious”, ”ingenious”, ”insightful”, ”intellectual”, ”intelligent”, ”knowledgeable”, ”leader”, ”logical”, ”meticulous”, ”organised”, ”original”, ”outspoken”, ”perceptive”, ”persistent”, ”power”, ”powerful”, ”practical”, ”proud”, ”rational”, ”realist”, ”recognition”, ”resilient”, ”resourceful”, ”serious”, ”sharp”, ”smart”, ”status”, ”strong”, ”superiority”, ”tough”, ”unique”, ”unwavering”, ”vigorous”

### 18.1.5 *Appearance Words*

”alluring”, ”voluptuous”, ”blushing”, ”homely”, ”plump”, ”sensual”, ”gorgeous”, ”slim”, ”bald”, ”athletic”, ”fashionable”, ”stout”, ”ugly”, ”muscular”, ”slender”, ”feeble”, ”handsome”, ”healthy”, ”attractive”, ”fat”, ”weak”, ”thin”, ”pretty”, ”beautiful”, ”strong”

### 18.1.6 *Violence Words:*

”homicide”, ”scourge”, ”unresolved”, ”devastate”, ”frustrate”, ”intolerant”, ”irritate”, ”uncontrolled”, ”uncorrected”, ”witness”, ”mele”, ”dehumanise”, ”nihilist”, ”thievery”, ”dissonance”, ”victimize”, ”puncher”, ”terrorize”, ”brutish”, ”wipeout”, ”behavior”, ”bloodthirsty”, ”destruct”, ”wrongdoer”, ”bestiality”, ”detest”, ”gunmen”, ”abduct”, ”abuse”, ”accomplice”, ”adversary”, ”affect”, ”affray”, ”afraid”, ”aggravate”, ”aggressive”, ”aggressor”, ”agitate”, ”allege”, ”altercation”, ”ambush”, ”ammunition”, ”anger”, ”angry”, ”anguish”, ”animosity”, ”annoy”, ”anxiety”, ”arm”, ”army”, ”arrest”, ”arson”, ”artillery”, ”assailant”, ”assault”, ”atomic”, ”attack”, ”bad”, ”bang”, ”barbaric”, ”battle”, ”beast”, ”beef”, ”bite”, ”bitter”, ”blast”, ”bleed”, ”blood”, ”bloodshed”, ”bloody”, ”bomb”, ”bomber”, ”brawl”, ”bruise”, ”brutal”, ”brute”, ”bullet”, ”bully”, ”burglary”, ”burn”, ”campaign”, ”casualty”, ”charge”, ”collision”, ”combat”, ”command”, ”complain”, ”conflict”, ”confront”, ”consequence”, ”convict”, ”corrupt”, ”crass”, ”craze”, ”crime”, ”criminal”, ”cruel”, ”crusade”, ”crush”, ”cry”, ”culprit”, ”cult”, ”cut”, ”damage”, ”danger”, ”dead”, ”death”, ”decease”, ”defeat”, ”defend”, ”demeaning”, ”demise”, ”demolish”, ”demolition”, ”denigrate”, ”deploy”, ”deprive”, ”despair”, ”destroy”, ”detention”, ”detriment”, ”die”, ”differ”, ”difficult”, ”disagree”, ”discontent”, ”disgrace”, ”disgust”, ”disillusioned”, ”displeasure”, ”dispute”, ”dissatisfaction”, ”distress”, ”domestic”, ”dread”, ”elicit”, ”encounter”, ”endanger”, ”enemy”, ”escalate”, ”evasive”, ”execute”, ”explode”, ”explosive”, ”extreme”, ”fatal”, ”fear”, ”felon”, ”felony”, ”fierce”, ”fight”, ”fire”, ”foe”, ”force”, ”fought”, ”fright”, ”fury”, ”gang”, ”gash”, ”gore”, ”grab”, ”grave”, ”grenade”, ”grief”, ”grievous”, ”guard”, ”gun”, ”gunman”, ”gunshot”, ”hack”, ”hamstring”, ”harass”, ”harm”, ”harsh”, ”hassle”, ”hate”, ”hatred”, ”hit”, ”horrific”, ”hostile”, ”human”, ”hurt”, ”hysterical”, ”imperialist”, ”imprison”, ”impulse”, ”incident”, ”incite”, ”increase”, ”inexplicable”, ”inflame”, ”inflict”,

”injure”, ”injury”, ”inmate”, ”intend”, ”intent”, ”intimidate”, ”invasive”, ”investigate”, ”irate”, ”ire”, ”issue”, ”jab”, ”jail”, ”jealousy”, ”jihad”, ”kick”, ”kidnap”, ”kill”, ”knife”, ”law”, ”loot”, ”lose”, ”loss”, ”machete”, ”mad”, ”maim”, ”mangle”, ”mayhem”, ”mean”, ”meaningless”, ”menace”, ”military”, ”mindless”, ”missile”, ”mistrust”, ”molest”, ”munition”, ”murder”, ”mutilate”, ”nuclear”, ”obliterate”, ”offence”, ”offend”, ”onslaught”, ”oppose”, ”opposite”, ”oppress”, ”oppressor”, ”outburst”, ”outrage”, ”pain”, ”passion”, ”peril”, ”perpetrate”, ”person”, ”pointless”, ”poison”, ”police”, ”power”, ”prejudice”, ”pressure”, ”prey”, ”prison”, ”protest”, ”provoke”, ”punch”, ”puncture”, ”push”, ”racism”, ”racist”, ”radical”, ”rage”, ”raid”, ”ram”, ”rampage”, ”rape”, ”rapist”, ”raze”, ”rebel”, ”red”, ”resent”, ”riot”, ”rip”, ”risk”, ”robbery”, ”roughshod”, ”round”, ”ruin”, ”ruthless”, ”savage”, ”scene”, ”scrape”, ”scream”, ”screech”, ”scuffle”, ”senseless”, ”sentence”, ”serious”, ”severed”, ”sexual”, ”shameless”, ”shoot”, ”shot”, ”shout”, ”showdown”, ”shrapnel”, ”shriek”, ”sicken”, ”slash”, ”slay”, ”slice”, ”smash”, ”snap”, ”snipe”, ”soldier”, ”spite”, ”spree”, ”stab”, ”strength”, ”stress”, ”strike”, ”struck”, ”struggle”, ”suffer”, ”suicide”, ”suspect”, ”swing”, ”target”, ”tear”, ”tense”, ”tension”, ”terror”, ”terrorist”, ”theft”, ”thief”, ”threat”, ”threaten”, ”thrust”, ”traffic”, ”trauma”, ”trial”, ”troop”, ”trouble”, ”undercut”, ”unease”, ”unhappy”, ”unpredictable”, ”unruly”, ”unsafe”, ”useless”, ”vandal”, ”vehement”, ”vicious”, ”victim”, ”violence”, ”violent”, ”war”, ”warfare”, ”warhead”, ”weapon”, ”weaponry”, ”wild”, ”worry”, ”worse”, ”worst”, ”wound”, ”wrath”, ”yell”

## 18.2 GENDER CLASSIFICATION LEXICON

### 18.2.1 *Male Words*

”man”, ”boy”, ”he”, ”father”, ”son”, ”guy”, ”male”, ”his”, ”himself”, ”grandpa”, ”grandpas”, ”grandson”, ”grandsons”, ”uncle”, ”husband”, ”boy”, ”brother”, ”dad”, ”dude”, ”fella”, ”gentleman”, ”gentlemen”,”men”, ”nephew”, ”nephews”, ”sir”, ”sirs”, ”lad”, ”mr”, ”daddy”, ”boys”, ”guys”, ”sons”, ”uncles”, ”sons”, ”misters”, ”mister”, ”daddies”, ”sons”, ”fellas”, ”stepfather”, ”stepfathers”, ”dads”

### 18.2.2 *Female Words*

”woman”, ”girl”, ”she”, ”mother”, ”daughter”, ”gal”, ”gals”, ”female”, ”her”, ”hers”, ”herself”, ”grandma”, ”grandmas”, ”granddaughter”, ”aunt”, ”wife”, ”wives”, ”sister”, ”sisters”, ”mum”,”mums”, ”gal”, ”granny”, ”lady”, ”women”, ”niece”, ”nieces”, ”ladies”, ”mrs”, ”ms”, ”mummy”, ”girls”, ”daughters”, ”aunts”, ”daughters”, ”misses”, ”missus”, ”grannies”, ”mummies”, ”stepsister”, ”sisters”, ”mummys”, ”stepmother”, ”stepmothers”

## 18.3 FEMALE COEFFICIENTS

| feature | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| headscarf | 18.43 | 11.68 | 17.82 | 14.29 | 5.72 | 9.56 | 10.34 | 7.26 | 7.07 | 11.18 | 11.75 | 4.70 | 3 |
| pregnant | 10.04 | 11.92 | 8.28 | 10.08 | 8.98 | 10.06 | 10.07 | 8.22 | 8.21 | 7.86 | 10.05 | 9.07 | 5 |
| hijab | 5.43 | 10.59 | 9.19 | 12.16 | 12.10 | 9.51 | 7.96 | 8.84 | 10.01 | 9.21 | 8.37 | 5.52 | 3 |
| breastfeed | 14.36 | 11.60 | 9.25 | 8.25 | 7.15 | 7.20 | 8.33 | 9.25 | 9.60 | 7.77 | 5.66 | 4.26 | 2 |
| legging | 14.96 | 6.89 | 15.85 | 9.52 | 8.76 | 8.14 | 7.67 | 9.60 | 4.90 | 8.11 | 5.37 | 1.96 | 0 |
| ovarian | 5.27 | 14.19 | 12.87 | 11.14 | 7.25 | 8.85 | 5.60 | 7.37 | 5.76 | 7.44 | 5.57 | 4.09 | 0 |
| bikini | 11.05 | 7.70 | 5.88 | 7.40 | 7.04 | 9.52 | 5.82 | 11.62 | 11.96 | 5.26 | 6.70 | 4.99 | 0 |
| pregnancy | 9.51 | 7.38 | 7.22 | 6.78 | 7.77 | 7.81 | 8.46 | 7.83 | 7.35 | 8.38 | 7.11 | 5.46 | 4 |
| sari | 8.44 | 10.93 | 7.22 | 13.26 | 4.28 | 10.30 | 11.13 | 6.04 | 4.06 | 8.69 | 9.79 | 0.63 | 0 |
| suffragette | 4.22 | 5.59 | 10.28 | 7.76 | 1.21 | 6.46 | 8.16 | 10.91 | 13.57 | 12.01 | 9.11 | 2.59 | 1 |
| fgm | 8.67 | 7.39 | 7.12 | 13.96 | 7.77 | 8.16 | 5.50 | 8.60 | 3.58 | 12.03 | 7.86 | 1.26 | 0 |
| actress | 8.08 | 8.78 | 7.40 | 8.98 | 7.41 | 7.63 | 5.52 | 6.40 | 6.34 | 6.52 | 7.80 | 5.42 | 3 |
| menopause | 4.07 | 8.68 | 4.61 | 3.12 | 3.51 | 11.01 | 7.27 | 9.49 | 9.90 | 11.48 | 7.20 | 5.58 | 2 |
| menstrual | 6.44 | 6.86 | 5.32 | 8.42 | 9.32 | 13.28 | 6.79 | 5.65 | 5.69 | 9.25 | 5.71 | 4.98 | 1 |
| heroine | 8.18 | 6.94 | 6.76 | 5.81 | 5.50 | 13.96 | 7.18 | 8.71 | 6.28 | 11.37 | 2.64 | 1.93 | 1 |
| childbearing | 12.75 | 8.71 | 6.45 | 9.48 | 8.26 | 5.54 | 4.13 | 7.39 | 5.18 | 6.24 | 7.62 | 2.81 | 0 |
| niqab | 7.15 | 9.95 | 15.23 | 5.94 | 7.65 | 6.39 | 8.50 | 6.34 | 6.50 | 3.87 | 5.30 | 1.43 | 0 |
| gaga | 4.90 | 8.97 | 4.90 | 8.77 | 7.21 | 10.14 | 4.99 | 6.01 | 5.96 | 9.54 | 9.54 | 3.14 | 0 |
| endometriosis | 9.16 | 2.41 | 7.62 | 3.46 | 10.81 | 6.70 | 6.63 | 9.85 | 6.55 | 9.69 | 6.50 | 3.12 | 1 |

Figure 39: 20 Words Most Associated to the Female Class

## 18.4 MALE COEFFICIENTS

| feature | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| prostate | -7.22 | -8.02 | -11.89 | -7.92 | -11.37 | -11.43 | -8.46 | -6.63 | -9.18 | -13.64 | -5.16 | -4.35 | -2.34 |
| knighthood | -9.52 | -12.45 | -7.19 | -4.74 | -2.81 | -13.21 | -7.02 | -7.52 | -10.00 | -1.29 | -10.35 | -2.93 | -1.75 |
| homily | -13.04 | -4.07 | -8.78 | -9.05 | -9.73 | -11.84 | -6.05 | -7.60 | -6.25 | -6.88 | -4.74 | -1.02 | -0.12 |
| fatherhood | -6.92 | -10.49 | -9.57 | -12.74 | -5.11 | -7.77 | -8.15 | -8.91 | -4.57 | -3.76 | -6.21 | -2.96 | -0.33 |
| cardinal | -12.80 | -3.47 | -5.05 | -8.73 | -14.28 | -11.60 | -4.01 | -5.09 | -2.57 | -4.86 | -6.71 | -2.04 | -0.58 |
| stubble | -3.18 | -6.53 | -4.38 | -10.96 | -5.20 | -10.68 | -9.20 | -4.09 | -8.93 | -8.78 | -6.10 | -3.20 | -0.43 |
| batsman | -6.22 | -7.75 | -10.51 | -11.39 | -10.70 | -9.17 | -7.69 | 2.10 | -1.60 | -7.53 | -6.34 | -2.35 | -0.26 |
| mujahideen | -6.91 | -10.15 | -9.17 | -9.28 | -9.65 | -7.69 | -5.74 | -5.72 | -3.14 | -5.33 | -5.49 | -0.42 | -0.04 |
| cumberbatch | -7.24 | -5.65 | -7.09 | -8.71 | -9.26 | -4.67 | -10.12 | -10.22 | -6.62 | -5.20 | -1.25 | -0.88 | -0.65 |
| zuckerberg | -10.08 | -10.01 | -9.22 | 1.12 | -8.48 | -5.00 | -10.70 | -1.82 | -3.55 | -7.85 | -8.41 | -2.26 | -0.37 |
| testicle | -12.76 | -1.73 | -6.42 | -12.90 | 0.40 | -2.80 | -6.53 | -5.91 | -3.25 | -10.28 | -9.79 | -3.24 | -0.80 |
| testicular | -10.20 | -1.29 | -4.89 | -2.33 | -8.78 | -9.92 | -6.17 | -6.49 | -8.85 | -9.24 | -3.53 | -1.51 | -0.31 |
| rmt | -5.75 | -9.91 | -7.49 | -4.28 | -10.54 | -2.70 | -10.14 | -4.96 | -6.51 | -5.41 | -4.60 | 0.49 | -0.65 |
| underpants | -7.66 | -3.57 | -11.93 | -1.56 | -8.40 | -6.48 | -9.77 | -9.44 | -7.05 | -3.34 | -0.97 | -0.66 | 0.07 |
| emir | -8.70 | -7.68 | -1.32 | -7.36 | -2.92 | -6.49 | -5.97 | -5.69 | -5.98 | -8.76 | -8.43 | -0.52 | -0.01 |
| affable | -8.07 | -2.84 | -1.76 | -11.08 | -8.34 | -9.41 | -7.16 | -5.58 | -4.32 | -5.34 | -2.89 | -1.79 | -0.24 |
| genial | -3.03 | -5.28 | -7.90 | -8.46 | -8.65 | -11.49 | -6.97 | -6.98 | -3.57 | -4.38 | -1.13 | -0.76 | -0.16 |
| binman | -6.38 | -9.55 | -7.58 | -6.10 | -6.23 | -4.00 | -4.94 | -9.56 | -3.85 | -3.24 | -6.47 | -0.48 | -0.14 |
| incrimination | -7.02 | -7.31 | -3.46 | -4.15 | -10.71 | -6.72 | -6.74 | -9.24 | -5.28 | -3.41 | -0.47 | -2.59 | -1.02 |

Figure 40: 20 Words Most Associated to the Male Class

## 18.5 EQUATIONS

### 18.5.1 *TF-IDF*

$$w_{i,j} = t_{i,j} \times log(\frac{N}{df_i}) \tag{1}$$

18.1: TF-IDF Formula

### 18.5.2 *Logistic Regression*

$$y_i = f(x_i\beta) + \varepsilon_i \tag{2}$$

18.2: Logistics Regression Formula

### 18.5.3 *Performance Metrics*

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \tag{3}$$

18.3: Accuracy Formula

*TP* = *true positive*, *FP* = *false positive*, *TN* = *true negative* and *FN* = *false negative*.

$$\text{Macro Average} = \frac{\sum_{i=j}^{n} \text{Accuracy}_i}{n} \tag{4}$$

18.4: Macro Accuracy Formula

$$\text{F1 Score Formula} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{5}$$

18.5: F1 Formula

$$\text{Macro F1 Score} = \frac{\sum_{i=j}^{n} \text{F1 Score}_i}{n} \tag{6}$$

18.6: Macro F1 Score