UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Innovation and Technology Management Curriculum

Devrim Nesipoglu

Comparison of toxicity among female and male active politicians in social media

Master's Thesis (20 ECTS)

Supervisor(s): Uku Kangur, Rajesh Sharma

Tartu 24 April 2024

Comparison of toxicity among female and male active politicians in social media

Abstract:

Toxicity analysis on social media is critical in understanding and addressing hateful behaviors and discourse. This master's thesis aims to comprehensively compare toxicity levels in online social media discourse among American male and female politicians. The study uses a multifaceted machine learning approach and natural language processing (NLP) techniques. Leveraging sentiment analysis, we measure the sentiment of posts and comments made by politicians while using models for toxicity detection to classify text as toxic or non-toxic. In addition, we separate the data into male and female categories, thus enabling a detailed comparison. Statistical analysis is then applied to assess and compare toxicity levels between the two groups, shedding light on possible gender-based differences in online discourse. Through visualization and interpretation of results, we aim to contribute to understanding toxicity patterns in social media on political communication and gender dynamics.

Keywords:

Hate speech, toxicity analysis, social media analysis, natural language processing, sentiment analysis, text mining

CERCS: P175, Informatics, systems theory

Naiste ja meeste aktiivsete poliitikute mürgisuse võrdlus sotsiaalmeedias

Lühikokkuvõte:

Toksilisuse analüüs sotsiaalmeedias on vaenu õhutava käitumise ja diskursuse mõistmisel kriitilise tähtsusega. Selle magistritöö eesmärk on võrrelda toksilisuse taset internetipõhises sotsiaalmeedia diskursuses aktiivsete Ameerika mees- ja naispoliitikute näitel. Uuringus kasutatakse mitmekülgset masinõppe lähenemist ja loomuliku keele töötlemise (NLP) tehnikaid. Töös mõõdetakse poliitikute postituste ja kommentaaride sentimenti ning kasutatakse toksilisuse tuvastamise mudeleid, mis klassifitseerivad teksti toksiliseks või mittetoksiliseks. Lisaks eraldatakse andmed meeste ja naiste kategooriatesse, võimaldades soopõhist toksilisuse võrdlust. Seejärel rakendatakse statistilist analüüsi, et hinnata ja võrrelda kahe rühma toksilisuse tasemeid, valgustades võimalikke soopõhiseid erinevusi veebidiskursuses. Tulemuste visualiseerimise ja tõlgendamise kaudu soovime aidata mõista toksilisuse mustreid sotsiaalmeedias poliitilise kommunikatsiooni ja soolise dünaamika osas.

Võtmesõnad:

Vihakõne, toksilisuse analüüs, sotsiaalmeedia analüüs, loomuliku keele töötlemine, sentimentide analüüs ja tekstikaevandamine.

CERCS: P175, Informaatika, süsteemiteooria

Table of Contents

1 Introduction

Political discourses have changed form with the increasingly effective use of social media in recent years. Politicians effectively use social media platforms to interact with the public through the internet and social media. The public can see posts shared by politicians and respond either instantaneously or retrospectively. In this way, they can interact with politicians without intermediaries. Compared to the era before the internet and social media, it is a big opportunity for both sides.

People's ability to express themselves freely has not only had its positive aspects, but it has also allowed individuals to speak toxically without hesitation. Toxic speech refers to any harmful, hurtful, or offensive communication. It covers various behaviors and language that can belittle or harm individuals, groups, or communities. This speech may include insults, harassment, bullying, or derogatory comments. It may appear online and offline in various forms of communication, such as social media, forums, messaging platforms, or personal interactions.

1.1. Background

In the political world, active male and female politicians often face online toxicity, which can significantly impact their public image (Stieglitz, 2012). Especially in the political arena, the issue has gained considerable attention (Liboiron, 2018). With the rise of social media platforms in recent years, political discourse has moved to a virtual arena where discussions, debates, and ideas are exchanged (Sobieraj, 2022). This case is especially true for female politicians, who often face identity-based hate and sexual intimidation, both online and offline (Dion, 2018). Online abuse and harassment have a significant impact on women's political voice and visibility (Sobieraj, 2022).

Social media platforms provide anonymity and mobility, making it easier for hateful behavior and speech to become more widespread, which poses significant challenges (Udanur, 2019). This study investigates the impact of toxic behaviors on political polarization, public opinion, and democratic processes, particularly within political interactions on social media platforms.

1.2. Problem Statement

The research will focus on analyzing toxicity among American politicians' posts and comments by respondents. The goal is to provide practical insights that can help improve understanding of the gender effect on toxicity speech, which can create useful insights into political science.

1.3. Contribution of the Thesis

By analyzing and understanding toxicity levels in online social media discourse directed at male and female politicians, this study contributes to existing knowledge comprehensively. Natural language processing approaches, including sentiment analysis, topic modeling, and toxicity detection, are used to analyze Facebook comments and posts. Comparing sentiment distribution, toxicity levels, and engagement metrics between male and female politicians can enlighten gender-based differences in online discourse and user interaction. Additionally, the study explores novel research questions related to sentiment analysis and toxicity detection in political communication on Facebook, contributing to a deeper understanding of online political discourse.

The thesis uses a comparative analysis approach, using data collected from Facebook, one of the most popular social media platforms worldwide, to examine toxicity patterns in comments directed at male and female politicians. Data analysis techniques include natural language processing, topic modeling, and comparative statistical methods.

Text classification is generally required for toxicity detection in speech. It involves extracting features from text data and using classification models to detect hate speech. The research involved collecting, cleaning, and analyzing data from Facebook, a social media platform. There are two research questions that are being addressed in this study:

RQ1: *How do the sentiments and discussion topics within the comments and posts directed at male and female politicians' Facebook posts differ?*
RQ2: *What are the differences in engagement between the male and female politicians' comments on Facebook?*

The thesis identifies significant disparities in the frequency and severity of toxic interactions experienced by male and female politicians with the help of research questions. Gender differences are observed in the context of toxic language used and the targeted nature of harassment, highlighting the need for targeted interventions to address online gender-based violence.

1.4. Organization of the Thesis

Structure of the Thesis:

The remainder of this thesis is organized in the following structure: an introduction, literature review, methodology, results, discussion, and conclusion. Each chapter is structured to provide a comprehensive overview of the research process, analysis, and results.

For this research work, there are seven (7) chapters.

- Chapter 1 is the Introduction, which provides a detailed background of the research work, explaining the reasons for the research work and the problem statement.
- Chapter 2 provides the Theoretical Background and is divided into three sections. The first section discusses related works on sentimental analysis of Hate speech data, and section two discusses related work.
- Chapter 3, Data Collection and Preprocessing, discusses in detail how data was collected and pre-processed to ensure it was clean enough for exploratory data analysis.
- Chapter 4 discusses the methodology, including natural language processing and machine learning algorithms.
- Chapter 5 discusses the results and the analysis of the two main research methods. Every step of the analysis is broken down here to understand different insights extracted from the data.
- Chapter 6 discussion
- Chapter 7 concludes this research work and examines future works for further analysis.

2 Theoretical Background

Detecting hate speech and toxicity encountered in social media, which has become a part of our lives with the great leap forward in technological developments in recent years, has been an exciting subject for researchers. The theoretical Background section provides an overview of existing research on hate speech detection, highlighting critical studies that evaluate various issues. By addressing this gap in the literature, the study seeks to provide insights into the unique challenges faced by American politicians, contribute to our understanding of gender dynamics in online political communication, and offer practical recommendations for fostering a healthier digital political discourse.

2.1. Hate Speech and Toxicity in Political Discourse

Toxicity in online social media discourse is commonly referred to as using language or behavior perceived as offensive, harmful, or disrespectful towards others. Extensive research has been conducted to define and recognize toxicity, with some studies providing broad descriptions that encompass any form of communication that belittles individuals or groups based on specific characteristics (Nockleby, 2002), while others offer more particular classifications that concentrate on hate speech, cyberbullying, or harassment.

The reason for choosing to examine toxicity in the context of political discussions on social media platforms is that it is a widespread issue that can have significant consequences for democratic engagement and political participation. Active female and male politicians in the United States have been chosen as a demographic group to provide relevant insights and practical recommendations for the country's current socio-political landscape.

Identifying hate speech in the text is challenging, even for humans. That is why it is crucial to establish clear definitions of hate speech before using machine learning to identify it. However, there is no single formal definition of hate speech; a widely accepted standard definition was proposed by Nockleby (2002): "Any form of communication that disparages a person or group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other defining trait."

The prevalence of toxicity encountered in comments on social media can lead to more toxic reactions due to sensitive topics such as racism and war (Salminen, 2020). Despite increasing research on hate speech and the topic receiving increasing attention, more studies are needed to understand and address this issue (Montero, 2022).

Examining toxicity in the context of social media and political discourse is essential due to its pervasive and detrimental impacts. Research studies highlight how critical it is to combat these negative impacts. Alshamrani (2020) emphasizes the connection between toxic behaviors and various news stories, particularly those about crime and religion.

Empirical analysis of hate speech datasets has been explored by numerous articles, including Arango, Pérez, and Poblete (2019), Paula Fortuna, and others. They focused on analyzing hate speech datasets to understand the prevalence and characteristics of hate speech online. Their study may have examined the language used in hate speech, the sources of hate speech data, and the effectiveness of different detection methods. They utilize natural language processing techniques (NLP), sentiment analysis to identify hate speech content, and machine learning algorithms for classification tasks.

Anjum and Katarya (2023) contribute to understanding toxicity in online communication. They use different methodologies and datasets. Paula Fortuna and Anjum with Katarya (2023) may have conducted a comprehensive survey to explore the state of the art in hate speech and toxicity detection in online social media. Their study likely involved analyzing various research papers, datasets, and methodologies used for hate speech detection.

By studying hate speech datasets and analyzing the detection methodologies, these researchers have contributed valuable insights to the ongoing efforts to combat hate speech and toxicity in online communication.

2.2. Gender Disparities in Hate Speech

Gender dynamics in online political communication are complex and multifaceted. Both Koc-Michalska (2019) and Hu (2020) highlight the impact of social media platforms on these dynamics, with Koc-Michalska (2019) explicitly noting the popularity of "mansplaining" in political discussions on Twitter. Hu (2020) further emphasizes Twitter's role in promoting the voices of marginalized groups regarding COVID-19 vaccines, including women. Maximova (2020) and New (2001) both explore gender differences in self-representation and participation in political communication, with Maximova (2020) finding that women tend to have less visible role models and less political engagement on Facebook.

Research shows that female politicians can be the target of hate speech and online harassment. Solovev (2022) found that women serving in the US Congress were more likely to receive hate speech on Twitter. Wilhelm (2018) emphasizes the gendered nature of moral judgments and more closely evaluates women's online behavior. Döring (2020) also found that female YouTubers are more likely to receive sexist and sexually offensive comments. This gender disparity in online harassment has significant consequences and creates a hostile environment for women in politics (Wagner, 2020).

Female politicians often face higher levels of toxicity and online abuse compared to male politicians across various social media platforms (Wagner, 2022; Rheault et al., 2019). Despite the toxicity, female politicians tend to receive more supportive engagement on social media, indicating lower toxicity levels in specific contexts (Samuel-Azran & Yarchi, 2023). There are explicit gender biases towards politicians on online platforms, with female politicians being subjected to more sexist comments, hate speech, and gender-based abuse (Marjanovic et al., 2022; Solovev & Pröllochs, 2022).

Ethnicity, appearance, and sexual history can significantly influence the type and intensity of online harassment female politicians face (Kužel et al., 2022; Esposito & Breeze, 2022). As the visibility and status of female politicians increase, so does the incidence of online incivility and harassment (Tromble & Koole, 2020). While toxicity levels vary, political affiliation may influence the nature of online harassment experienced by politicians, with some political parties receiving more negativity regardless of gender (Fichman & McClelland, 2021).

Gender plays a significant role in how online harassment towards politicians is perceived, with female politicians often facing heightened toxicity due to potential misogyny (Phillips et al., 2023). Both female and male politicians exhibit similar behaviors on social media, but gender differences are observed in how positive interactions are received (Just & Crigler, 2014).

Table 1 shows a list of various studies regarding gender disparities and summarizes their findings:

Title, Author, and Citation	Main Findings
Gender Differences in Abuse: The Case of Dutch Politicians on Twitter by K Chandra Shekar (2023)	Contrary to expectations, female politicians on Twitter scored lower in toxicity compared to males, as per the study on Dutch politicians' online abuse. Male politicians received higher levels of toxicity on Twitter than female politicians, except for threats. Female ethnic minority politicians faced the highest threat levels. (Shekar, 2023)
The "gender affinity effect" behind female politicians' social media support: Facebook civil talk during Israel's 2021 elections by Tal Samuel-Azran +1 more (2023)	Female politicians receive more supportive engagement on social media than male politicians, indicating lower toxicity towards them. The study focuses on the effect of gender affinity on online political discourse. (Samuel-Azran & Yarchi, 2023)
Gender, Digital Toxicity, and Political Voice Online by Sarah Sobieraj (2020)	Female politicians face higher levels of digital toxicity, including identity-based hate and sexual intimidation, impacting their political voice online compared to male politicians. (Sobieraj, 2020)
Quantifying gender biases towards politicians on Reddit by Marjanovic, S., Stańczak, K., & Augenstein, I. (2022)	Female politicians face more nominal and lexical biases, with comments often focusing on personal attributes. Toxicity levels differ, with males receiving more coverage but similar comment lengths. (Marjanovic et al., 2022)
How Women Politicians of Fiji are Treated on Facebook by Rasťo Kužel +3 more (2022)	Female politicians on Facebook face more sexist comments compared to male politicians. Male politicians receive four times more problematic content, but female politicians are targeted with comments on personal traits rather than politics. (Kužel et al., 2022)
Gender and politics in a digitalised world: Investigating online hostility against UK female MPs by Eleonora Esposito +1 more (2022)	The study found varying levels of online hostility towards UK female MPs, with some receiving more toxic comments related to appearance, sexual history, and violence compared to male MPs. (Esposito & Breeze, 2022)
The impact of gender and political affiliation on trolling by Pnina Fichman +1 more (2021)	Female politicians experience more trolling on social media than male politicians, as indicated by the research findings on gender impact in political trolling on Twitter. (Fichman & McClelland, 2021)
Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada by Angelia Wagner (2022)	Female politicians face heightened online toxicity due to potential misogyny, as highlighted in the paper. Gender plays a significant role in the perception of online harassment towards politicians. (Wagner, 2022)
Running While Female: Using AI to Track how Twitter Commentary Disadvantages Women in the 2020 U.S. Primaries by Sarah Oates +3 more (2019)	Female politicians face more toxicity on social media compared to male politicians, with attacks on character and identity being prominent, reflecting biases seen in traditional media coverage. (Oates et al., 2019)
As the Tweet, so the Reply?: Gender Bias in Digital Communication with Politicians by Armin Mertens +3 more (2019)	A study analyzing digital interactions during the German federal elections 2017 revealed that female politicians face more gender-based toxicity on social media than male politicians. (Mertens et al., 2019)
Politicians in the line of fire: Incivility and the treatment of women on social media by Ludovic Rheault +2 more (2019)	According to the study, female politicians face more incivility on social media than male politicians, significantly as their visibility and status increase. (Rheault et al., 2019)

Table 1. Comparison of Studies on Gender Differences in Online Toxicity Towards Politicians

2.3. Social Media Platforms and Political Engagement

Several studies have investigated hate speech and the toxicity of political discourse, especially towards politicians. Alkomah (2022) emphasizes the complexity of hate speech and the need for more reliable datasets to detect it. Agarwal (2021) presents a case study of British MPs on Twitter, revealing that hate speech is more common during peak periods and is often directed at ethnic minorities or MPs who hold positions in the government. Paz (2020) and Gracia-Calandín (2023) highlight the importance of interdisciplinary approaches and the need for ethical reflection in combating hate speech; Gracia-Calandín (2023) specifically calls for a systematic evaluation of proposals to combat hate speech.

Various methods have been studied to analyze toxicity on social networks. Malmasi (2017) and Alkomah (2022) highlight the difficulties distinguishing hate speech from blasphemy; Alkomah emphasizes the need to achieve consistent results on different types of hate speech. Garg (2022) and Risch (2020) focus on the biases and limitations of existing methods; Garg proposes a taxonomy of unconscious bias, and Risch discusses the need for a detailed taxonomy of feedback. These studies provide a comprehensive overview of the complexities and potential solutions in toxicity testing.

2.4. Methodologies for Analyzing Toxicity

Traditional methodologies for detecting hate speech and toxicity of politicians on social media involve the application of machine learning and deep learning techniques. Researchers have proposed various approaches utilizing these methods to identify hate speech automatically on online social media platforms (Meng et al., 2022; Awal et al., 2021; Cao et al., 2020).

Studies have shown the effectiveness of models like Random Forest and BERT in detecting hate speech content (Alkomah et al., 2022). The gravity of the issue has prompted both social media platforms and academic researchers to develop and propose traditional machine learning and deep learning solutions for automatic hate speech detection (Awal et al., 2023).

These methodologies aim to classify harmful comments and prevent the dissemination of toxic content on social media networks (Luu et al., 2022). The detection of hate speech on social media faces challenges such as imbalanced datasets and selecting appropriate models and feature analysis methods (Romim et al., 2021). Automated hate speech identification is crucial due to the vast amount of content generated on social media, making manual moderation impractical (Elzayady et al., 2023). Additionally, transfer learning approaches based on pre-trained language models like BERT have been introduced to automatically detect hateful speech in social media content (Han et al., 2021).

Research into the toxicity of online comments has shown that the choice of English as the language can influence the identification and translation of toxic language in studies. Kobellarz (2022) finds it best to keep comments in their original language, while Costa-jussà (2021) notes the prevalence of additional toxicity in low-resource languages during machine translation. This toxicity was attributed to translation errors, hallucinations, and unstable translations. These results show that the language used can significantly influence the identification and translation of toxic language and that additional toxicity is a particular concern in low-resource languages.

2.4.1. Text Mining and Machine Learning Algorithms

Text mining involves preprocessing textual data to extract meaningful information and patterns, including data cleaning, standardization, and tokenization to prepare the text for further analysis. This method is essential in analyzing hate speech and toxicity in textual data. Text mining techniques, such as sentiment analysis and topic modeling, automatically detect hate speech. These methods have effectively classified text as hate speech, aiding in identifying offensive language and sentiments.

Various text mining and machine learning algorithms have been applied to detect toxicity. Helma & Kazius (2006) emphasized the importance of these tools in deriving toxicity estimates and explainable models from toxicity data. These studies highlight the potential of text mining and machine learning algorithms for toxicity detection.

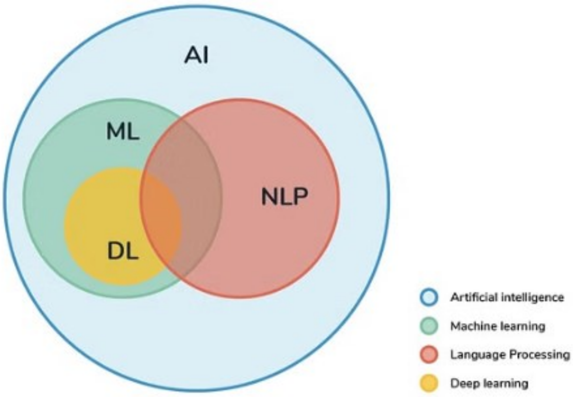
Previous research has emphasized the importance of lexicons in hate speech detection, text mining, and sentiment analysis (Luu et al., 2022). Additionally, studies have shown that Text Mining (TM), Information Retrieval (IR), or Natural Language Processing (NLP) techniques are more effective for hate speech identification than Keyword-based or Rule-based mining approaches (Qureshi & Sabih, 2021).

2.4.1.1. Natural Language Processing (NLP)

As one of the computational techniques, Natural language processing (NLP) is utilized to analyze, understand and extract insights from textual data efficiently. It is an interdisciplinary field at the crossing point of Computer Science, Artificial Intelligence, and Linguistics (Hassan, A. 2018) (Bacco et al., 2022). Especially with the revolutionary developments in the field of machine learning (ML) and artificial intelligence (AI), data mining with NLP, more effective use of text recognition, and predictive modeling have been significantly improved (Karhade et al., 2022).

The diagram (Figure 1) given below illustrates the intersection and overlap between four critical areas of technology: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). Each of these fields plays a crucial role in analyzing social media data.

Figure 1. Intersection of AI, ML, DL, and NLP in Social Media Analysis



2.4.1.2. Sentiment Analysis

Sentiment analysis (SA) uses natural language processing (NLP) algorithms to analyze the emotion conveyed in text content. The process typically entails classifying text into positive, negative, or neutral sentiments based on the emotional tone expressed. Furthermore, within political analysis, SA is a crucial tool for deciphering public opinion concerning political issues and candidates (Ali, 2023).

- **Lexicon-Based Sentiment Analysis with Vader or TextBlob**

Vader and TextBlob are two widely used tools for sentiment analysis, and each offers unique features and capabilities. Vader, a dictionary and rule-based sentiment analysis tool, is victorious in capturing the sentiment polarity of the text, especially in social media content where informal language and expressions are ordinary. TextBlob, on the other hand, provides a simplified interface for everyday NLP tasks, including sentiment analysis, through its intuitive API and pre-trained models.

In our analysis pipeline, Vader and TextBlob effectively measure sentiment expressed in social media posts and comments toward political figures. Using these tools, this study aims to capture the emotional nuances in politicians' comments and responses.

- **NLTK (Natural Language Toolkit):**

In the field of Natural Language Processing (NLP), NLTK (Natural Language Toolkit) is a widely utilized library. It provides various functions necessary for text analysis and processing. Some specific functionalities of NLTK include tokenization, resource allocation, lemmatization, and sentiment analysis (Jongeling et al., 2017).

Tokenization involves breaking text into individual words or sentences, a fundamental step in NLP tasks (Ames & Havens, 2021). The stemming and lemmatization of words helps normalize and analyze text (Jongeling et al., 2017). An analysis of sentiment or emotion in a text can be useful for various purposes, such as understanding customer feedback or social media sentiment (Jongeling et al., 2017).

2.4.2. Topic Modeling

Social media data can be analyzed using topic modeling to identify underlying themes and patterns. In the context of toxicity analysis, social media topics can provide insight into prevalent themes related to toxicity, such as hate speech, cyberbullying, and offensive language. It is possible to interpret these topics based on the associated words and phrases, which provides a deeper understanding of the nature and extent of toxic content in the dataset (Salminen et al., 2020).

BERTopic and Latent Dirichlet Allocation (LDA) are commonly used by researchers to identify topics in the data automatically (Egger & Yu, 2022). They are advanced tools for analyzing social media data, especially in the context of toxicity analysis.. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a widely used topic modeling technique applied to various domains, including hate speech and toxicity detection. LDA works by representing each word in a corpus as a mixture of underlying topics, allowing for identifying themes within a text collection (Sear et al., 2022). This method has been particularly effective in identifying hate topics within text associated with online communities that promote hate (Sear et al., 2022). By modeling words as combinations of topics, the LDA method has been chosen as a topic modeling technique in this research to capture the underlying toxicity or hate concepts in Facebook comments or posts of politicians (Ombui et al., 2021).

In the context of hate speech detection, LDA has been instrumental in uncovering specific manifestations of hate speech, such as racism, xenophobia, sexism, and misogyny (Chiril et al., 2021). Additionally, LDA has been used to track the evolution of online hate topics, providing insights into the changing landscape of hate speech on digital platforms (Sear et al., 2022). Furthermore, LDA is effective in dealing with sparse and short data, making it a valuable tool for analyzing microblogs and other text datasets (Wang et al., 2015).

LDA is a powerful tool for detecting hate speech and toxicity. It allows researchers to uncover underlying hate concepts, track hate topics online, and analyze text data efficiently.

2.4.3. Toxicity classifiers

Detecting and reducing toxic content is essential in social media analysis to ensure a safe and constructive online environment. Therefore, the methodology of this study includes the following advanced techniques for toxicity classification that aim to identify and analyze toxic language patterns in the collected data. Perspective API is accessed through Google, on the other hand Bert and Detoxify is accessed through Hugging Face platform. Hugging Face provides pre-trained models and libraries for various NLP tasks, including toxicity detection. 'Transformers' is a library provided by Hugging Face that offers interfaces for working with various pre-trained transformer models, including BERT. Transformers libraries in Python are used to load pre-trained models and perform toxicity classification.

2.4.3.1. The Perspective API

The Perspective API, developed by Google's technology incubator Jigsaw, is a widely utilized toxicity classifier employed by various online platforms to detect and filter out toxic comments, aiming to maintain a safe online environment (Reichert, 2020). Integrating the Perspective API into an analysis pipeline involves utilizing machine learning to automatically identify toxic language, thereby enhancing toxicity detection processes' efficiency.

The *'**Perspective API*** scores posts and comments based on perceived toxicity. The scoring system generates a probability score between 0 and 1. If the score has higher values, it shows a greater likelihood that the comment is toxic. Perspective API sends HTTP requests directly to its REST API endpoints provided by Google. I used Python libraries to simplify interacting with the API.

2.4.3.2. Detoxify

Detoxify is a comment detection library introduced by Hanu and the Unitary team in 2020, which utilizes Hugging Face's transformers to identify inappropriate or harmful text online (Chhablani, 2021). Technically, it utilizes the capabilities of the Hugging Face library and is made available through Hugging Face's platform**.** This library has been applied in toxicity analysis for this thesis to detect harmful, or bad words within online content. The identification of toxic behavior, including hateful comments or toxicity, is crucial in various online platforms and social media interactions (Salminen et al., 2020).

Hugging Face is a well-known open-source library that provides a variety of pre-trained models and general-purpose architectures for natural language processing (NLP) tasks (Boukabous & Azizi, 2021). This library played an essential role in making NLP open source, making it more accessible to researchers and end users (Boukabous & Azizi, 2021).

In the context of hate toxicity analysis, Detoxify leverages transformer-based language models such as BERT to adapt to hate speech detection tasks (Luu et al., 2022). These models are designed to help identify online hate speech by processing toxic and hateful text.

2.4.3.3. BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) is a language representation model introduced by (Devlin et al., 2018). Unlike previous models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by considering both left and proper context in all layers (Devlin, 2018). This bidirectional approach allows BERT to capture a more comprehensive understanding of the context in which words appear.

Since our dataset for analysis is not labeled, the BERT method can be applied to detect and classify toxicity in the ingredient. BERT's bidirectional nature and deep contextual understanding make it a powerful tool for analyzing sensitivity and detecting toxicity. For instance, Fan et al. (2021) applied BERT to detect and classify toxicity in social media content related to the UK Brexit. The authors leveraged BERT's capabilities to understand the nuances of language and classify sentiment categories effectively.

Furthermore, BERT has been used in hate speech detection (Mozafari et al., 2019), offensive language detection (Isaksen & Gambäck, 2020), and even in detecting complex sensitive sentences. Its effectiveness in various NLP tasks, including sentiment analysis and emotion detection, has been widely acknowledged (Sosea & Caragea, 2021).

3 Data and Preprocessing

3.1. Data Collection and Instruments

The research for hate speech detection utilizes a dataset collected from Facebook by web crawling as a large-scale text. The dataset consists of the posts and comments made on Facebook by politicians who play an active role in the United States of America and the responses and reactions they received. Names are chosen systematically according to their roles in the US government and their active use of Facebook. Male and female politicians' numbers are selected equally. The gathered datasets provide a comprehensive view of public engagement with political content on social media.

Position	Politician	Position Definition
Joe Biden	President	The head of state and government is responsible for the overall federal government administration.
Kamala Harris	Vice President	The second-highest-ranking official in the U.S. government, crucial in supporting the President. Marco Rubio
Marco Rubio	U.S. Senate	One of the two chambers of the U.S. Congress, with specific legislative and advisory responsibilities.
Tommy Tuberville, Alex Padilla, Katie Boyd Britt, Kevin McCarthy, Steny Hoyer, Elise Stefanik, Laphonza Butler	U.S. House of Representatives	The other chamber of the U.S. Congress focuses on population representation.

Table 2: Politicians names and positions that are used in the study

3.2. The Rationale for Dataset Selection for American politicians

America is considered one of the leading countries where citizens can freely express themselves, primarily in English, one of the most widely spoken languages globally. According to statistics, English is the most commonly used language on social networks (58.8%) and on more than (50.0%) of websites (Omran et al., 2023). This motivates the language selection for this study.

In the United States, citizens are proud of freedom of expression and opportunities for citizen participation (Reines, 2016). It is one of the foundations of democracy that citizens can express their opinions without fear of censorship (Gunawan et al., 2021). This freedom of self-expression is vital for the functioning of a democratic society, as it allows individuals to express their concerns, demands, and opinions. Social media platforms have further strengthened the opportunity for individuals to express themselves freely and make their voices heard globally.

Another important reason for choosing this country for analysis is that American politicians actively use Facebook. According to the Reuters Institute Digital News Report 2023, Facebook is still the leading global social media and messaging platform. The report highlights Facebook's continued dominance as the primary source for news consumption across various demographics. The rich dataset obtained from Facebook provides a comprehensive view of their interactions and public responses. Facebook's unique feature allows users to express themselves with relative freedom and allows for a broader range of political discourse. This freedom of expression is necessary to investigate the toxicity and gender dynamics and to contribute to a subtle analysis.

The study of toxicity and gender differences in political discourse on Facebook is critical because it sheds light on the complexities inherent in online political interactions. It offers insights into the nature and impact of toxic discourse, and understanding how gender dynamics emerge in these interactions is crucial for promoting inclusive and respectful digital political spaces.

3.3. Data Scraping Method

The primary scraping process involved in this study is creating an instance of the Facebook scraper class with the specified parameters and scraping posts' content from the given (Facebook) page. The politicians were selected systematically from an equal number of men and women among the highest-ranking politicians currently serving in America. It was preferred that they actively use social media and Facebook.

The obtained data is then meticulously preprocessed, including decoding text columns and renaming columns for consistency. Next, the code iterates through the post IDs, retrieves comments for each post, and extracts relevant information such as comment ID, commenter details, comment text, timestamp, and reactions. Finally, the scraped data is saved in CSV format, one for posts and another for comments, with filenames based on the specified page name.

The Facebook scraping method utilized a Python script with the following packages:

- facebook_page_scraper, pandas, Facebook scraper

A Python script was used to scrape data from Facebook pages using the mentioned packages. The Facebook scraper module facilitated the extraction of posts and comments from Facebook pages, gathering data for analysis. The page name variable was configured to target specific pages, such as "Joe Biden," allowing the script to collect data from the chosen lawmakers. The gathered datasets include variables such as post content, postdate, comments, and commenter details.

3.4. Data Preparing, cleaning, and pre-processing

The following preprocessing techniques given in (Table 3) are used for cleaning and preparing data. Politicians' Facebook comments and posts are collected into CSV files and transformed into Excel files. The files were merged into one Excel file using the ‘post_id’ of politicians as the primary key by the database program. After having one file for each politician, male and female politicians’ files are put into two separate files with the name male and female datasets to obtain gender-based datasets. After the preprocessing methods mentioned below were performed, the data number decreased from 594892 to 551424 in the female dataset and 46990 to 46535in the male dataset.

Facebook users write in daily language, often including uncommon or noisy characters; thus, analyzing the data without text cleaning and preprocessing is challenging. To overcome this problem, data must be cleaned and prepared for analysis. The following steps outline the detailed text preprocessing methodologies. These steps ensure that the data is cleaned and prepared for the subsequent sentiment analysis, hate word extraction, and Perspective API scoring, enabling a comprehensive evaluation of toxicity levels among active American politicians on Facebook.

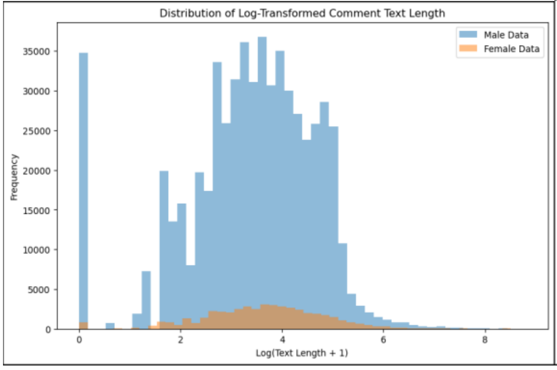
Preprocessing Technique	Description	Computational Techniques
Remove Special Characters	Eliminate non-alphanumeric characters such as punctuation marks and symbols.	String manipulation, Regular Expressions
Handle Missing Values	Identify and address missing entries within the dataset.	Data Imputation, Statistical Techniques
Eliminate Stop-Words	Remove common words such as "the", "is", and "are" from the text data.	NLP Libraries (e.g., NLTK, spaCy)
Strip HTML Tags	Extract text content and remove HTML markup elements.	String manipulation, Regular Expressions
Filter Out URLs	Detects and removes URLs or hyperlinks present in the text.	Regular Expressions
Remove Mentions	Identify and eliminate references to usernames or handles (e.g., "@username").	Regular Expressions
Purge Emojis and Emoticons	Detects and removes emoticons and emojis from the text.	Regular Expressions
Filter Out Hashtags	Identify and exclude hashtags or topic identifiers (e.g., "#topic").	Regular Expressions
Remove Extra Whitespaces	Identify and eliminate redundant whitespace characters.	String manipulation
Handle Numbers	Detect and address numerical characters within the text.	Regular Expressions
Remove Duplicate Entries	Identify and eliminate duplicate rows from the dataset.	Pandas (Python library)
Remove Commenter Names	Detects and removes commenter names or identifiers from text.	NLP Libraries (e.g., spaCy)

Table 3. List of Preprocessing that applied

The figures below show the situations before and after preprocessing and removing outliers at the data.

Figure 2a: Distribution of Log-Transformed Comment Text Length

Figure 2b: Excluding Outliers from Comment Text Length



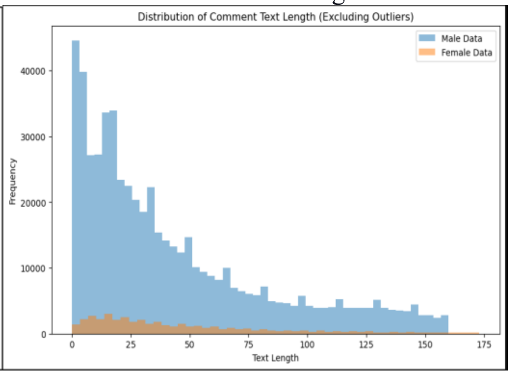
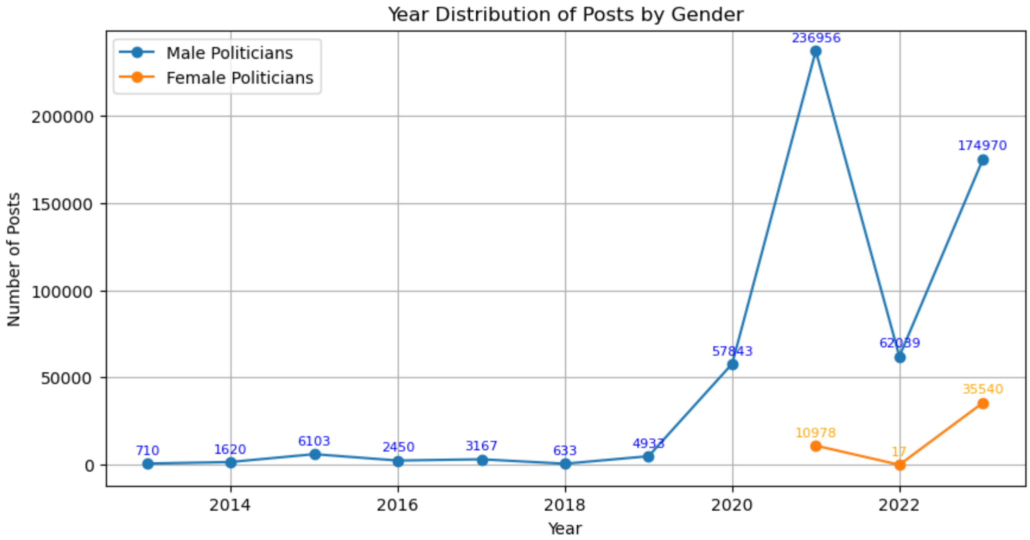


Figure 2a illustrates the distribution of log-transformed comment text lengths for both male and female datasets. The x-axis represents the log-transformed text length (with an added constant of 1 to avoid a logarithm of zero), while the y-axis indicates the frequency of occurrence. The histogram shows how comment lengths are distributed across the dataset, allowing for comparison between male and female comments. Figure 2b shows after excluding outliers from the comment text length distribution in both male and female datasets.

Table 4 and Figure 3 represents the distribution of posts over the years for both male and female datasets.

Figure 3: Distribution of posts over the years for both male and female datasets



Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
Female									10978	17	35540	**46535**
Male	710	1620	6103	2450	3167	633	4933	57843	236956	62039	174970	**551424**

Table 4. Distribution of posts over the years for both male and female datasets

Based on the numbers shown in Table 4 and Figure 3:

- The total number of posts for female data is 46,535, while for male data, it is 551,424.
- Female data shows a considerable increase in posts from 2021 to 2023, with the highest number recorded in 2021 (10,978 posts) and 2023 (35,540 posts).
- Male data also shows a significant increase in posts from 2020 to 2023, with the highest number of posts recorded in 2021 (236,956) and 2023 (174,970).

3.5. Exploratory Data Analysis / Basic Statistics

The dataset includes prominent figures from both genders holding various roles in the U.S. Congress. Table 5 presents the distribution of posts and comment numbers among politicians after preprocessing of included in the study.

For male politicians, the dataset encompasses 5,637 posts and 598,480 comments, with each politician contributing differently to this total. U.S. Senator Marco Rubio stands out among them with the highest posts and comments in the study dataset. In contrast, the dataset for female politicians comprises 3,787 posts and 47,717 comments, with Elise Stefanik, US Senator, contributing significantly.

Name	Gender	Role	Amount of "posts" per politician	Amount of "comments" per politician	Number of posts and comments on the preprocessed datasets
Joe Biden	Male	President	844	148,778	126,079
Marco Rubio	Male	US – Senator	998	242,287	219,062
Kevin Mc Carthy	Male	US - House of Representatives	795	35,859	31,735
Steny Hoyer	Male	US - House of Representatives	1,000	11,837	10,945
Alex Padilla	Male	US – Senator	1,000	13,053	1,277
Tommy Tuberville	Male	US – Senator	1,000	146,666	122,824
Total numbers of **Male** Dataset			5,637	598,480	511,922
Kamala Harris	Female	Vice President	1,000	10,522	9,148
Senator Katie Boyd Brit	Female	US – Senator	330	1,608	1,515
Laphonza Butler	Female	US – Senator	571	129	107
Elise Stefanik	Female	US – Senator	1,000	29,736	19,154
Alexandria Ocasio-Cortez	Female	US – Senator	886	5,722	5,405
Total numbers of **Female** Dataset			3,787	47,717	35,329

Table 5. List of politicians and distribution of posts and comments

Tables 6a & 6b provide the summary of statistics for male and female politicians based on engagement metrics. These metrics are shares, likes, loves, wow reactions, cares, sad reactions, angry reactions, haha reactions, total reactions count, comments count.

Metric	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Shares	563.22	923.9	0	56	270	628	7800
Likes	3884.41	5661.14	0	737	2000	4800	45000
Loves	592.51	1145.51	0	0	168	592	9800
Wow	1.64	17.34	0	0	0	0	309
Cares	28.5	215.5	0	0	0	0	2600
Sad	52.31	337.63	0	0	0	0	4700
Angry	109.17	360.73	0	0	0	0	3700
Haha	267.15	696.27	0	0	0	393	7800
Reactions Count	4935.69	6828.59	0	1084	2772	6063	45000
Comments	1640.1	1885.71	0	300	1000	2200	10000

Table 6a. Summary Statistics for Male Politicians Facebook Posts

Metric	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
Shares	153.86	297.47	0	6	38	145	3400
Likes	884.18	1363.56	0	80	458	1100	14000
Loves	168.81	435.94	0	0	23	187	6100
Wow	0.3	3.41	0	0	0	0	51
Cares	9.63	67.54	0	0	0	0	884
Sad	16.83	98.88	0	0	0	0	1100
Angry	47.7	192.31	0	0	0	0	1200
Haha	81.07	105.75	0	0	37	119	486
Reactions Count	1208.52	1818.31	0	132	650	1665	20100
Comments	523.57	580.48	0	185	315	643	10000

Table 6b. Summary Statistics for Female Politicians Facebook Posts

	Male Leader Comments	Female Leader Comments
Reactions Distribution		
Average Shares	23.8	135.45
Average Likes	197.9	634.29
Average Loves	31.1	83.88
Average Cares	5.97	14.03
Comment Length		
Before Preprocessing (words)	18.46	22.23
After Preprocessing (words)	8.11	10.35

Table 7: Comparison of Comments on Male and Female Leaders: Reactions Distribution and Comment Length

These results given in Table 7 underscore significant differences in the reactions and comment length between comments directed at male and female leaders, emphasizing the importance of our research.

Reactions Distribution:

- Shares: On average, comments on posts by female leaders receive significantly more shares (135.45) than those by male leaders (23.8). This implies that content posted by female leaders tends to generate more engagement in sharing.
- Likes: Similarly, comments on posts by female leaders receive substantially more likes (634.29) than those by male leaders (197.9), indicating a higher level of positive engagement.
- Loves and Cares: While the difference is less pronounced, comments on posts by female leaders also receive more love and care on average than on posts by male leaders.

Comment Length:

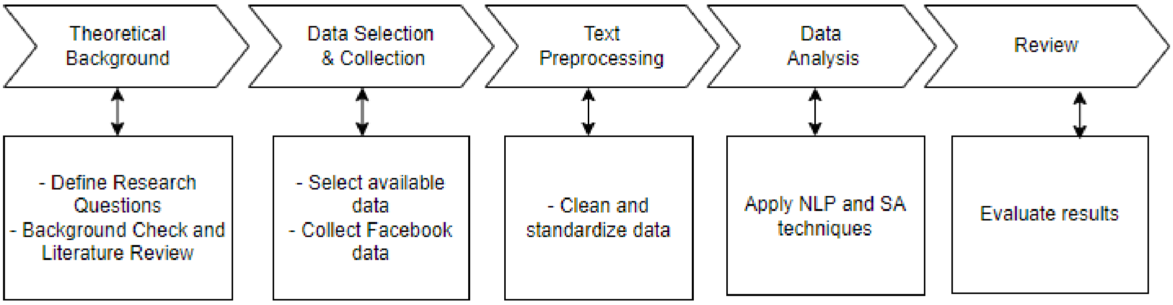
- Before Preprocessing: Comments on posts by female leaders have a slightly longer average length (22.23 words) than on posts by male leaders (18.46 words) before any preprocessing steps are applied.
- After Preprocessing: Cleanup or normalization steps, comments on posts by female leaders still tend to be slightly longer on average (10.35 words) compared to those on posts by male leaders (8.11 words).

Overall, these findings suggest that comments on posts by female leaders tend to attract more engagement in shares, likes, loves, and cares, indicating a higher level of positive interaction and support than comments on posts by male leaders. Additionally, comments on posts by female leaders tend to be slightly longer on average, both before and after preprocessing, which may suggest a higher level of detail or engagement with the content.

4 Methodology

The methodology section of this study outlines the tools and techniques employed in analyzing Facebook activities of prominent American political figures. The methodology framework in this study is shown in Figure 4 in five main parts. After deciding on research questions, some of the politicians are selected from the Congress web site. Their Facebook posts and comments are collected and preprocessed. In order to examine these research questions, data analysis steps are given in the following sections:

Figure 4. Methodological framework for analyzing collected politicians' posts and comments


Theoretical Background
Data Selection & Collection
Text Preprocessing
Data Analysis
Review
- Define Research Questions
- Background Check and Literature Review
- Select available data
- Collect Facebook data
- Clean and standardize data
Apply NLP and SA techniques
Evaluate results

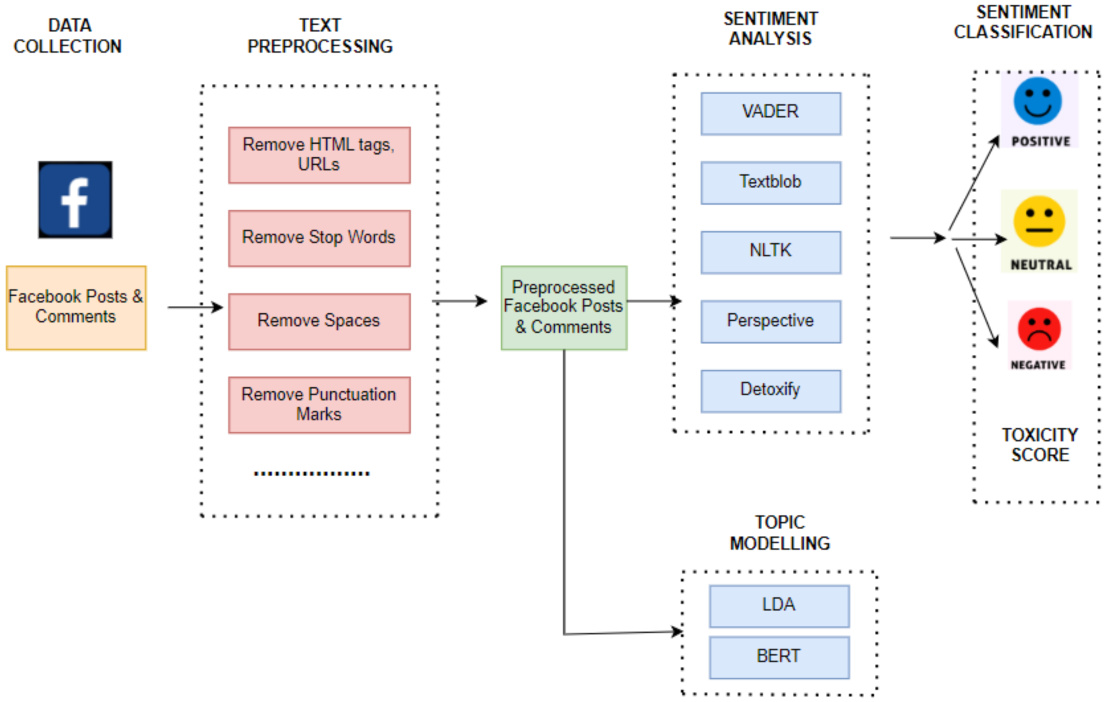
Before delving into the specific methodologies, providing a broad overview of the analysis pipeline is essential. The methodologies used in this thesis given in Table 8 summarize the steps. The pipeline includes:

Methodology	Description
Sentiment Analysis	The sentiment polarity of comments directed at male and female politicians on Facebook was analyzed using sentiment analysis tools, TextBlob and NLTK's Vader.
	Sentiment scores were calculated using these tools to assess the overall sentiment of the comments, categorizing them as positive, negative, or neutral.
Topic Modeling	Latent Dirichlet Allocation (LDA) was employed as a powerful topic modeling technique .
	It was applied to the document-term matrix to extract topics providing a clear understanding of the themes and issues discussed in the comments/posts.
Toxicity Analysis	Examined toxicity within social media discussions directed at female and male politicians using various toxicity classifiers, including the Perspective API, Detoxify, and BERT (Bidirectional Encoder Representations from Transformers).
	These classifiers were applied to assess the level of toxicity in comments/posts and to compare toxicity levels between male and female politicians.
Engagement Analysis	Conducted analysis to understand users' engagement with content posted by male and female politicians.
	Engagement metrics such as likes, shares, and reactions were calculated and compared between male and female politicians to evaluate audience engagement and interest.

Table 8: Methodologies Applied in This Study

A combination of sentiment analysis, topic modeling, toxicity analysis, and engagement analysis methodologies (Figure 5) are utilized in the Thesis to analyze social media interactions involving male and female politicians on Facebook.

Figure 5. Computational Methods Applied in This Study



The algorithm is used to determine sentiment labels and it is based on word level sentiment scores. Iteration through each comment in the dataset, ensures a precise calculation of the toxicity score for each comment. It keeps track of the total sentiment score and the count of toxic comments encountered. After processing all comments, if toxic comments are present (count_toxic_comments is not 0), it calculates the average toxicity score. Based on the average sentiment score it determines the sentiment:

- If the average score is less than 0.5, it considers the sentiment 'Positive.'
- If the average score exceeds 0.5, it considers the sentiment 'Negative.'
- The sentiment is considered neutral if the average score is NaN or equal to 0.5.

Conclusively, the algorithm provides a clear and definitive output, representing the determined sentiment. The top 10 positive, negative, and neutral words are identified based on their frequency and average word level sentiment score. Positive words generally express favorable sentiments, negative words express unfavorable sentiments, and neutral words are contextually neutral or lack sufficient toxicity data.

5. Results

This section summarizes the findings based on the analysis of the research questions. The thesis analyzes how Facebook users interact with political figures, particularly analyzing sentiment, engagement, and the statistical distribution of their reactions and comments. The framework integrates theories and methodologies from Natural Language Processing (NLP), primarily focusing on techniques such as topic modeling, hate word extraction, and sentiment analysis. Table 9 shows the results in categories.

Category	Description
Sentiment Analysis	Significant differences observed in sentiment distribution between comments directed at male and female politicians.
	Male politician comments exhibit a broader range of sentiment scores, suggesting polarization, while female politician comments display a more balanced sentiment distribution.
	Female politicians receive positive sentiments about leadership and integrity, while negative comments directed at female politicians often convey mistrust and skepticism.
	Male politicians are accused of criminal behavior, corruption, and scandal, while positive sentiments emphasize leadership and accountability.
Toxicity Examination	Male politician comments tend to have slightly higher toxicity levels than those directed at female politicians.
	Male politician comments receive slightly more positive sentiment scores than female politician comments.
Engagement Analysis	Engagement metrics analyzed include likes, shares, and reactions.
	Female politicians receive fewer engagement metrics on average compared to male politicians.
	Positive associations, albeit weak correlations, were observed between the number of negative remarks and "sad" reactions for both male and female politicians.
	Engagement metrics such as shares, likes, and reactions are positively associated with the number of comments, indicating a potential correlation between engagement levels and audience engagement.
Gender-Specific Differences	Minor gender-specific differences in the correlation between negative comments received by politicians and user engagement metrics were observed.
	Weak positive correlations were observed between negative comments and "sad" reactions for both male and female politicians, suggesting other factors may influence user participation in political posts.

Table 9: Categorization of Results

5.1. Sentiment Distribution

This section analyzes the sentiment distribution of comments directed at male and female politicians and provides findings.

Calculating Word-Level Sentiment Scores

"Word-level sentiment score" calculates sentiment polarity or emotional tone assigned to each word in a text. In sentiment analysis, words are often classified as positive, negative, or neutral based on their emotional connotations. Calculating word-level sentiment scores involves assigning a sentiment value to each word in a text based on its emotional meaning. These scores can then be aggregated or analyzed in more detail to understand the text's overall sentiment or to identify sentiment patterns across different texts.

The following steps were undertaken to calculate the sentiment score for each word:

- Datasets for male and female politicians' FB posts and comments were collected and preprocessed to have cleaned text.
- The sentiment analysis functions were applied to the sentiment scores of the comments.
- The frequency of occurrence and the associated sentiment scores were tracked for each unique word in the comments.
- The average word-level score for each word was calculated by computing the mean of all associated sentiment scores.
-

Sentiment Analysis (SA) of Facebook Comments

Sentiment Analysis involves analyzing the emotions in text content and categorizing them as positive, negative, or neutral, providing valuable insights.***TextBlob's*** sentiment scores provide a polarity score indicating the sentiment of the text. The '***nltk_sentiment_analysis***' function utilizes the NLTK library's ***Vader*** module to perform sentiment analysis on the cleaned text data. It calculates a sentiment score using a predefined lexicon of words with assigned sentiment scores. The sentiment score indicates the text's overall sentiment polarity (positive, negative, or neutral).

Methodology:

Using Text Blob and NLTK Vader sentiment analysis tools, sentiment scores were calculated and then applied to compare Facebook comments of male and female politicians to assess the toxicity of comments. The process involves the following steps:

Two approaches were utilized for sentiment analysis:

TextBlob Sentiment Analysis: The 'TextBlob' library was used to analyze the sentiment polarity of each comment. Sentiment polarity ranges from -1 to 1 (negative to positive).

NLTK's Vader Sentiment Analysis: The 'SentimentIntensityAnalyzer' from NLTK (Natural Language Toolkit) was used to compute the compound sentiment score for each comment. A compound score represents the overall sentiment of the text, ranging from -1 (extremely negative) to 1 (extremely positive).

Findings:

Sentiment analysis results show significant differences in the sentiment distribution between comments toward male and female politicians (see Table 10).

In particular, the sentiment distribution for comments directed at male politicians shows a broader range of sentiment scores across Text Blob and NLTK's Vader categories. This suggests a more diverse and polarized political interaction involving male politicians.

On the other hand, the sentiment distribution for comments towards female politicians displays a more balanced representation of positive, negative, and neutral sentiments across the Vader and Text Blob categories (see Table 10).

Gender	Sentiment Category	Vader	Vader %	Text blob	Text blob %
Male	Positive	226467	41.07%	266733	48.37%
Male	Neutral	165129	29.95%	174395	31.63%
Male	Negative	159828	28.98%	110296	20.00%
Total		**551424**	**100 %**	**551424**	**100 %**
Female	Positive	17660	37.96%	20705	44.51%
Female	Neutral	16164	34.74%	15231	32.74%
Female	Negative	12698	27.29%	10586	22.75%
Total		**46522**	**100 %**	**46522**	**100 %**

Table 10: Sentiment Distribution by Gender of Politicians' Comment Sections

Table 10 provides an insightful breakdown of average sentiment scores based on gender for both politicians' posts and comments on social media platforms. Male politicians tend to exhibit higher sentiment scores in their posts, with an average Vader score of 12.1% and a Text Blob score of 8.9%. However, regarding comments, the sentiment scores for males decrease, with an average Vader score of 6.8% and a Text Blob score of 5.1%. Conversely, female politicians display a contrasting pattern, with lower sentiment scores in their posts (Vader: 8.0%, Text Blob: 12.0%) but higher sentiment scores in comments (Vader: 1.2%, Text Blob: 3.6%). These findings shed light on the nuanced emotional dynamics of online interactions among politicians of different genders.

Gender	Sentiment Category	Vader %	Text blob %
Male	Average sentiment score for posts	12.1%	8.9%
Male	Average sentiment score for comments	6.8%	5.1%
Female	Average sentiment score for posts	8.0%	12.0%
Female	Average sentiment score for comments	1.2%	3.6%

Table 11: Average Sentiment Scores by Gender of Politicians' Posts and Comments

Comparison of the top 10 negative and positive words associated with comments

This section compares the top 10 negative and positive words associated with comments directed at female and male politicians. A Python random generator selected 2000 records equally from the male and female datasets and then used them in the analysis.

Our text processing method involves breaking down comments into individual words, a process known as tokenization. We then remove common words and punctuation and convert all words to lowercase for consistency. The frequency of each word is calculated using a Python Counter, and the results are stored in a Pandas Data frame. We also calculate the average toxicity score for each word across all comments containing it. Finally, we assign a sentiment label to each word based on its word-level sentiment score, categorizing them as 'Positive,' 'Negative,' or 'Neutral'.

- Words Associated with Female or Male Politicians: Analysis of prevalent negativity and positive sentiments, with examples (see Table 12).

Negative Words (Female Politicians)	Sentiment Score	Positive Words (Female Politicians)	Toxicity Score
Treachery	-0.9981	Privilege	1.00E-04
Betrayal	-0.9981	Leadership	0.000193333
Colluded	-0.9981	Integrity	0.000125
Slanderous	-0.9981	Empowerment	0.000142857
Perjury	-0.9981	Progress	0.000457704
Falsify	-0.9981	Advocacy	0.000648437
Treason	-0.9981	Equality	0.000585882
Defamatory	-0.9981	Inclusion	0.000457143
Defiance	-0.9981	Resilience	0.000457143
Accusations	-0.9981	Compassion	0.000595
Negative Words (Male Politicians)	**Sentiment Score**	**Positive Words (Male Politicians)**	**Toxicity Score**
Pedophiles	-0.9995	Leadership	9.35E-07
Treason	-0.9995	Integrity	2.54E-05
Racine	-0.9995	Accountability	3.86E-05
Seduction	-0.9995	Transparency	2.00E-05
Betrayal	-0.9981	Dedication	4.00E-05
Corruption	-0.9981	Commitment	4.00E-05
Scandal	-0.9981	Resilience	4.62E-05
Fraud	-0.9981	Empathy	7.69E-05
Collusion	-0.9981	Vision	0.000114286
Misconduct	-0.9981	Service	0.000125

Table 12. Top 10 Most Meaningful Negative and Positive Words for Female and Male Politicians

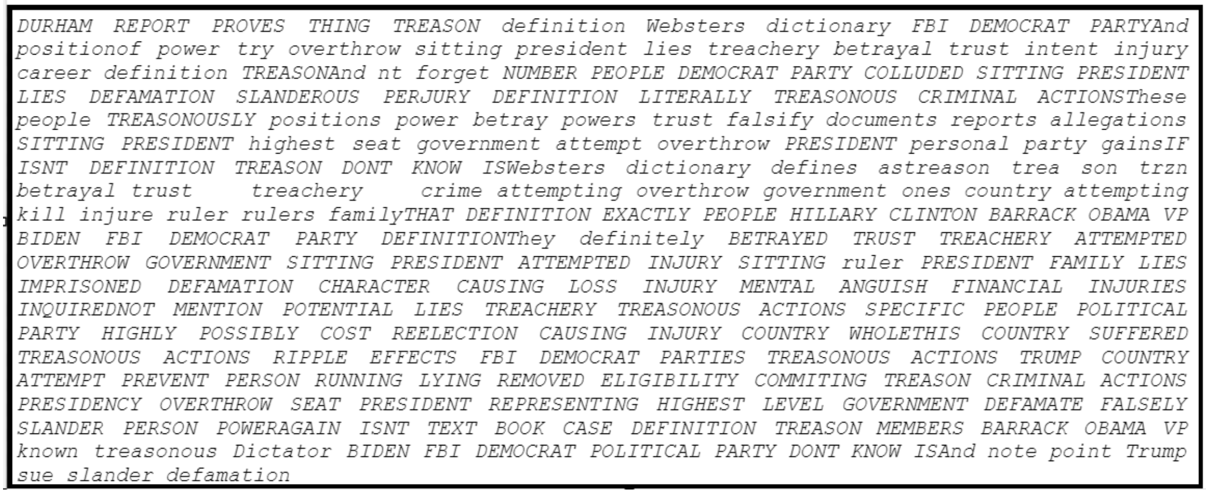
Discussion surrounding ***female*** politicians reveals a prevalent negativity, as evidenced by frequently using terms like 'treachery,' 'betrayal,' and 'colluded.' These words reflect a narrative steeped in distrust, deception, and ethical ambiguity, suggesting a challenging discourse environment for female political figures. Conversely, comments directed at female politicians often convey positive sentiments, with terms such as "leadership," "integrity," and "privilege" dominating the discourse. These words underscore the recognition of strong leadership qualities, ethical conduct, and acknowledgment of privilege within the discussion surrounding female politicians.

These top 10 words with posts are characterized by similar negativity, with words like "pedophiles," "treason," and "seduction" featuring prominently. These terms hint at severe allegations of criminal behavior, betrayal, and moral impropriety directed at male political figures.

On the other hand, positive sentiments directed at male politicians are often centered around terms like "leadership," "integrity," and "accountability." These words underscore the acknowledgment of desirable leadership qualities, ethical standards, and a commitment to public service within discussions surrounding male politicians.

Sample analysis of one word from the top 10 words

In-depth analysis of one word chosen from the top 10 negative words then analyzed , sentiment score and content analysis is provided as a result..



DURHAM REPORT PROVES THING TREASON definition Websters dictionary FBI DEMOCRAT PARTYAnd positionof power try overthrow sitting president lies treachery betrayal trust intent injury career definition TREASONAnd nt forget NUMBER PEOPLE DEMOCRAT PARTY COLLUDED SITTING PRESIDENT LIES DEFAMATION SLANDEROUS PERJURY DEFINITION LITERALLY TREASONOUS CRIMINAL ACTIONSThese people TREASONOUSLY positions power betray powers trust falsify documents reports allegations SITTING PRESIDENT highest seat government attempt overthrow PRESIDENT personal party gainsIF ISNT DEFINITION TREASON DONT KNOW ISWebsters dictionary defines astreason trea son trzn betrayal trust treachery crime attempting overthrow government ones country attempting kill injure ruler rulers familyTHAT DEFINITION EXACTLY PEOPLE HILLARY CLINTON BARRACK OBAMA VP BIDEN FBI DEMOCRAT PARTY DEFINITIONThey definitely BETRAYED TRUST TREACHERY ATTEMPTED OVERTHROW GOVERNMENT SITTING PRESIDENT ATTEMPTED INJURY SITTING ruler PRESIDENT FAMILY LIES IMPRISONED DEFAMATION CHARACTER CAUSING LOSS INJURY MENTAL ANGUISH FINANCIAL INJURIES INQUIREDNOT MENTION POTENTIAL LIES TREACHERY TREASONOUS ACTIONS SPECIFIC PEOPLE POLITICAL PARTY HIGHLY POSSIBLY COST REELECTION CAUSING INJURY COUNTRY WHOLETHIS COUNTRY SUFFERED TREASONOUS ACTIONS RIPPLE EFFECTS FBI DEMOCRAT PARTIES TREASONOUS ACTIONS TRUMP COUNTRY ATTEMPT PREVENT PERSON RUNNING LYING REMOVED ELIGIBILITY COMMITING TREASON CRIMINAL ACTIONS PRESIDENCY OVERTHROW SEAT PRESIDENT REPRESENTING HIGHEST LEVEL GOVERNMENT DEFAMATE FALSELY SLANDER PERSON POWERAGAIN ISNT TEXT BOOK CASE DEFINITION TREASON MEMBERS BARRACK OBAMA VP known treasonous Dictator BIDEN FBI DEMOCRAT POLITICAL PARTY DONT KNOW ISAnd note point Trump sue slander defamation

Figure 6: Word “Treachery” taken from the dataset comments for analysis

Analysis of Results are as follows:

Analysis of an example negative Word: TREACHERY

Sentiment Score: -0.9981

Content: Exposing Political Deception: A Deep Dive into Allegations of Treachery and Betrayal

The word "treachery" is highlighted in discussions related to the Durham Report, where it is used to accuse individuals associated with the Democratic Party of engaging in deceitful and disloyal behavior. The language portrays these actions as a betrayal of trust and an attempt to cause political and personal harm to the president and their family. This example showcases the word "treachery" being used to denote deceitful and disloyal actions attributed to specific individuals within the Democratic Party.

Comparative Analysis

Comparing the top 10 most negative and positive words associated with discussions directed at female and male politicians reveals nuanced differences in the nature and tone of online discourse. While both genders face criticism and scrutiny, the specific accusations and perceptions vary, reflecting distinct societal expectations, biases, and stereotypes.

Female politicians appear more frequently associated with accusations of betrayal, collusion, and treachery, suggesting a prevalent narrative of mistrust and skepticism surrounding their actions and motives. Conversely, discussions about male politicians often center around allegations of criminal behavior, corruption, and scandal, highlighting perceptions of moral lapses and ethical violations.

On the positive side, female and male politicians are praised for leadership, integrity, and dedication. However, the specific attributes emphasized in discussions about each gender may reflect underlying gender norms and expectations, with female politicians being lauded for their resilience and empowerment. In contrast, male politicians are commended for their accountability and transparency.

In conclusion, the comparison of toxicity among female and male active politicians in social media underscores the complex interplay of gender dynamics, societal perceptions, and political discourse within online platforms. Understanding these nuances is essential for fostering policies between politicians and the public.

	Toxicity Models			Sentiment Analysis Models	
Gender	**Perspective**	**BERT**	**Detoxify**	**TextBlob**	**Vader**
Male	0.0388	0.5613	0.1356	0.0407	0.056
Female	0.0208	0.5583	0.1399	0.0293	0.0218

Table 13: Comparison of the toxicity and sentiment average scores between male and female politicians

Toxicity Models: According to Perspective API, male politicians receive higher average toxicity scores (0.0388) than female politicians (0.0208). There is a negligible difference in average BERT toxicity scores between male (0.5613) and female (0.5583) politicians. Male politicians have slightly lower average Detoxify toxicity scores (0.1356) than female politicians (0.1399).

Sentiment Scores: Using TextBlob sentiment analysis, male politicians receive slightly higher average sentiment scores (0.0407) than female politicians (0.0293). Similarly, male politicians have higher average sentiment scores (0.056) than female politicians (0.0218) using VADER sentiment analysis.

Interpretation:

- The toxicity scores suggest that comments and posts directed at male politicians tend to have slightly higher toxicity levels than those directed at female politicians, especially according to Perspective API.
- However, the difference in toxicity scores between genders is relatively small, indicating that toxicity levels are generally comparable.
- Regarding sentiment, comments and posts directed at male politicians tend to have slightly more positive sentiments than those directed at female politicians, regardless of the sentiment analysis method used.

These findings, while insightful, only scratch the surface. They suggest that there may be differences like discussions and sentiments expressed towards male and female politicians on Facebook. However, a deeper understanding of the underlying factors contributing to these differences requires further qualitative analysis of the content and context of the comments and posts.

Analysis of Top 10 Comments with the Highest Sentiment Score

This analysis examines the top 10 comments with the highest sentiment scores extracted from our sample data (see Table 14). These comments provide valuable insights into the prevalent sentiments expressed in online discourse related to political topics.

Top 10 Positive Comments	**Dataset by gender**	**Sentimen t Score**
BIDEN WONT RETURNING WHITE HOUSESo glad Americans smartened	female	0.199041
bad Palestinian students unemployment rate apartheid regime	female	0.196723
boy obiden right Difference trump charges politically motivated obiden real criminal demoNcrats know beat trump communist regimes imprison opposition tried assassinate RFK Jr yesterday beat kill HEIL STALIN	female	0.196593
cackling Lifetime Racist Biden Turned Free Safe America Violent Crime Ridden World Sanctuary Cesspool	female	0.194605
heaps coals wicked souls	male	0.190022
Horrible criminal cancer society Yes	male	0.189634
admin plan enslave use immigrants army American ppl	female	0.188133
careless	female	0.187715
Israel Gaza antiSemite	female	0.181982
forgot giggle giggle haaahaahhhahh oh ohh giggle giggle hahahahha like smell yellow busses giggle giggle	female	0.17751
Top 10 Negative Comments	**Dataset by gender**	**Sentimen tScore**
FKH CACKLING IDIOT	female	0.997214
democrats communists quit supporting enslaved party slavery look dumb ass communist democrat idiots	male	0.996771
want abortions kill children year screaming rid gun save children bigger idiot Pedophile	female	0.996692
United Nation Antonio fuck wife pregnant wife bitch secretary sex work Antonia secretary eat ass eat mother poop bitch General Secretary United Nations Secretary split mouth bitch nose snooze mouth drink taste snooze split General Secretary time brutally	male	0.996491
effortless Republicans House Senate seatsAmericans stupid stupid	male	0.996455
America loses idiots	female	0.996239
Friggin idiot	female	0.995884
FUCK BOT fucking piece bribing SHIT	male	0.995849
wo taking pictures middle loser people smiling like stupid	male	0.995791
stoned STUPID	male	0.995755
Top 10 Neutral Comments	**Dataset by gender**	**Toxicity Score**
allah dead King	male	0.499988
Enemies freedom committed act war Country Text President speech Nation Muslim Extremist Terrorists attack United States America	male	0.494646
American speak nation world scammer	female	0.491571
intruders come home willa single shot multi shot gun better	female	0.491076
dead sound like Christianity	male	0.481228
intent murdering babies disgrace	female	0.476177
goddamned church government	male	0.47254
faced phony	male	0.457504
Americans died covid republicans incompetence	male	0.441838
Yemen America fake	male	0.430406

Table 14: Top 10 highest sentiment scores by distribution and emotion

5.2. Applying Topic Modelling (LDA)

The methodology uses computational techniques such as text processing and topic modeling (LDA) to separately identify discussion topics within posts and comments directed toward male and female politicians. Extracting insights from textual data and the distribution of issues will give insights into themes like toxicity, hate speech, or negativity. By applying these techniques, the study aims to identify topics discussed by male and female politicians on Facebook and analyze the sentiments expressed in their comments/posts. Comparing sentiment distribution allows for understanding potential differences in sentiment expression between male and female politicians across various topics.

The text data is first converted into a document-term matrix using the „Count Vectorizer," which tokenizes the text and builds a known word vocabulary. LDA is then applied to the document-term matrix to identify topics and associated word distributions.

Topic Modeling (LDA) Findings for Posts

Topics for Male Politicians' Posts:

The identified topics for male politicians include themes related to serving the people, the need for change, commitment, and national issues. These topics suggest that male politicians often discuss public service, societal needs, and national concerns in their posts.

Topics for Female Politicians' Posts:

Female politicians' topics revolve around making a difference, women's empowerment, policy initiatives, and announcements. These topics indicate that female politicians frequently address issues related to empowerment, policy advocacy, and announcements of initiatives or campaigns.

Topic Modelling (LDA) Findings for Comments

Topics for Male Politicians' Comments:

The topics discovered in the male dataset encompass various political issues, including voting, party affiliations, beliefs, and discussions about specific politicians like Trump and Biden. Additionally, topics related to gratitude ("thank"), religious references ("god"), and personal expressions ("happy birthday") were identified.

Topics for Male Politicians Comments:

Similarly, the topics in the female dataset cover political subjects such as voting, party affiliations, specific politicians (e.g., Biden, Trump), and beliefs. In addition to political topics, expressions of gratitude ("thank"), religious references ("god"), and personal expressions ("happy birthday") were also present.

Overall, the topics discussed in toxic comments directed toward male and female politicians on social media largely overlap. Both datasets contain discussions on political issues, expressions of gratitude, religious references, and personal expressions. However, further analysis may reveal subtle differences or patterns specific to each dataset.

Comparison of Top Comments per Topic in Male and Female Datasets

The following analysis compares the top comments from each topic within both the male and female datasets to understand the nature of toxicity and sentiment expression in comments directed toward male and female politicians.

1. The topics represent themes or subjects discussed in the comments directed at male and female politicians. For example, some of the topics for comments directed at male politicians include "Trump and country," "Democrats and military," "Voting and senators," "Happiness and money," and "Water and military bases."
2. The sentiment score used here is from TextBlob. TextBlob's sentiment analysis module assigns polarity scores to text, indicating the sentiment as positive, negative, or neutral. The polarity score ranges from -1 (negative) to 1 (positive).
3. These sentiment scores and associated topics are for comments directed at male and female politicians. The analysis focuses on understanding the sentiment expressed in these comments and identifying the prevalent topics discussed.

Gender	Topic Name	Average Sentiment	Top Words
Male	Politics and Governance	0.051	trump, country, world, help, vote
Male	Social Issues and Welfare	-0.036	country, democrats, military, amen, coach
Male	Economic Policies and Trade	0.031	voted, time, vote, senator, need
Male	Healthcare and Environment	0.155	happy, birthday, friend, way, money
Male	Foreign Affairs and Security	-0.025	water, bases, military, maybe, thank
Female	Politics and Governance	0.045	guns, crime, senator, war, percent
Female	Social Issues and Welfare	-0.034	need, trump, want, gop, gun
Female	Economic Policies and Trade	0.043	thank, trump, let, want, government
Female	Healthcare and Environment	0.008	american, going, think, way, money
Female	Foreign Affairs and Security	0.079	day, time, classified, party, vote

Table 15: Sentiment analysis of gendered discourse on various topics using TextBlob

1. **Accuracy of Sentiment Scores**: The sentiment scores calculated using TextBlob vary across topics and genders. They range from negative to positive, indicating a mix of sentiments in the comments directed at male and female politicians.

2. **Consistency of Topics**: The topics identified based on the top words associated with each topic appear to align with different aspects of political discourse. For example, topics like "Politics and Governance," "Social Issues and Welfare," "Economic Policies and Trade," "Healthcare and Environment," and "Foreign Affairs and Security" cover a broad spectrum of topics and genders.

3. **Variation Across Genders**: There are differences in sentiment scores and topics between comments directed at male and female politicians. For instance, in the "Politics and Governance" topic, comments directed at male politicians seem to focus on words like "Trump," "country," and "vote," while comments directed at female politicians focus on words like "guns," "crime," and "senator."

5.3. Engagement Analysis

This section aims to answer the second research question using the Engagement Analysis.

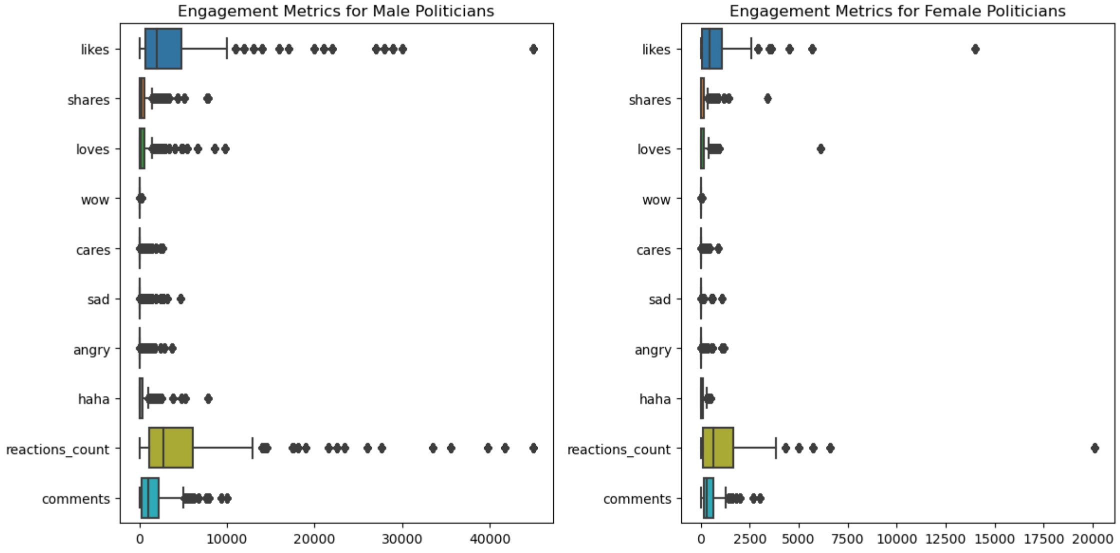
RQ2: *What are the differences in engagement between the male and female politicians' comments on Facebook?*

RQ2 focuses on how sentiment influences users' interactions with political information and examines the connection between user engagement. Study aims to find gender-specific variations and connections between negative reviews and other engagement measures.

Engagement Analysis assesses how social media users typically engage with content on social media platforms. It analyzes metrics such as likes, shares, comments, reactions, and views to understand audience engagement and interest in content. This type of analysis aims to gain insight into audience behavior, preferences, and interactions with content, which can be used in various areas, such as audience targeting and campaign optimization.

Reactions like 'shares,' 'like,' 'loves,' 'wow,' 'caring,' 'sad,' 'angry,' and 'ha-ha' are quantified as user engagement metrics in the collected data (Figure 7). The datasets for men and women are analyzed independently.

Figure 7: Engagement Metrics for Male and Female Politicians



In the dataset collected for this study, engagement metrics such as likes, shares, and reactions received by male and female politicians are calculated as an average or total and aggregated together for each gender group. The data was then analyzed to understand which gender received more engagement on social media and how audience engagement varied by gender.

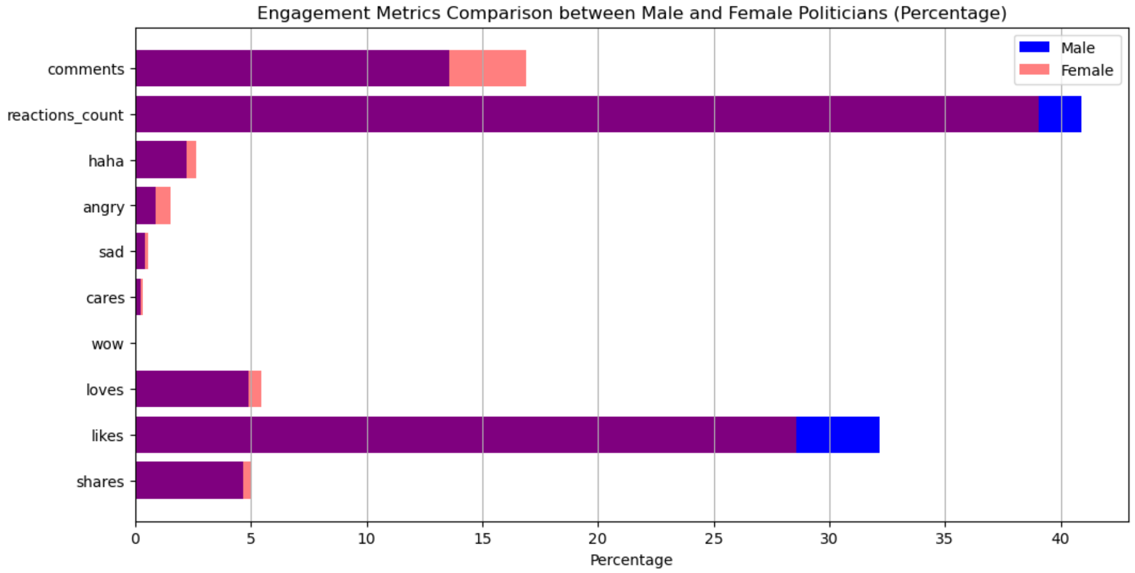
Engagement analysis results are as follows:

- Female politicians receive an average of 884.18 likes, 153.86 shares, and 1208.52 reactions to their posts in 48000 records.

- Male politicians, an average of 3884.41 likes, 563.22 shares, and 4935.69 reactions to their posts around half million records.

Comparing the engagement metrics of male and female politicians by percentage can provide insight into the analysis (Figure 7). There are no significant differences in metrics between genders. The biggest difference is seen in the comments. Comments from women outnumber comments from male politicians by approximately five percent.

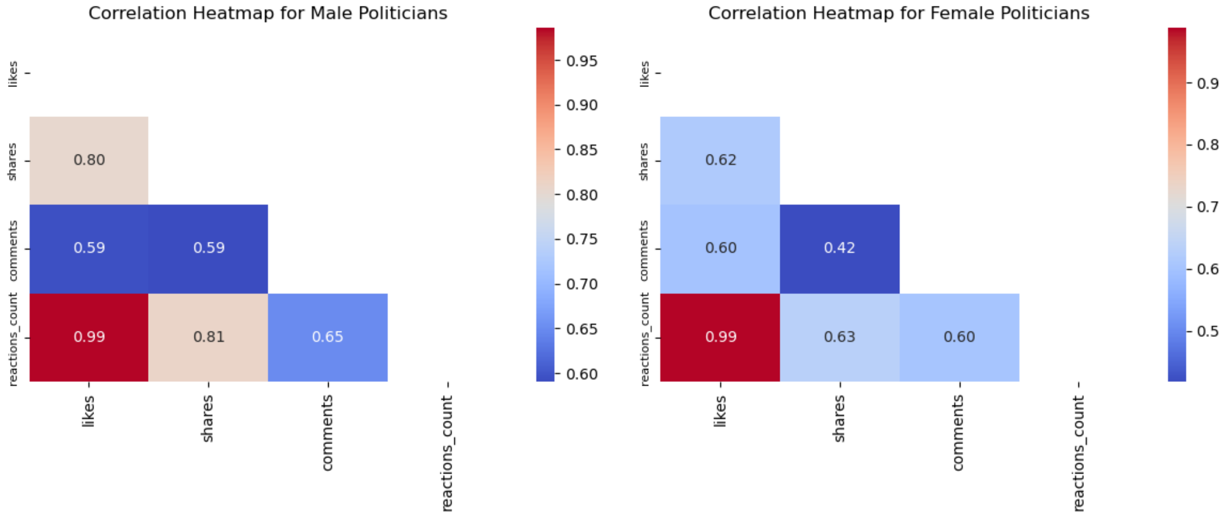
Figure 8: Engagement Comparison between Male and Female Metrics Politicians



Engagement metrics analysis of the politician’s posts

The correlation heatmaps shown in Figure 8 provide information of the relationships between different engagement metrics for male and female politicians on social media platforms. The strong positive correlations observed between likes and overall reactions indicate that posts receiving more likes tend to generate higher overall reaction engagement. Similarly, the positive correlations between likes and comments and between likes and shares suggest that posts with more likes also tend to attract more comments and shares. Additionally, the correlations between shares and reaction counts and between shares and comments further highlight the impact of shares on overall engagement and interaction.

Figure 9. Correlation Heatmap for Male and Female Politicians Posts



Both the male and female datasets show similar patterns of engagement and interaction on Facebook (Figure 8). Posts by both genders receive high levels of likes and overall reactions, indicating significant engagement from their respective audiences. However, there are slight variations in how these engagements manifest. For instance, posts by male politicians tend to receive slightly higher shares (0.18 on average) than those by female politicians, suggesting a potential difference in content dissemination strategies. Additionally, while both genders receive comments on their posts, female politicians' posts tend to attract a slightly higher number of comments relative to shares compared to male politicians (0.17 on average). These nuances in engagement patterns may play a role in understanding the dynamics of toxicity and the dissemination of toxic comments across genders on social media.

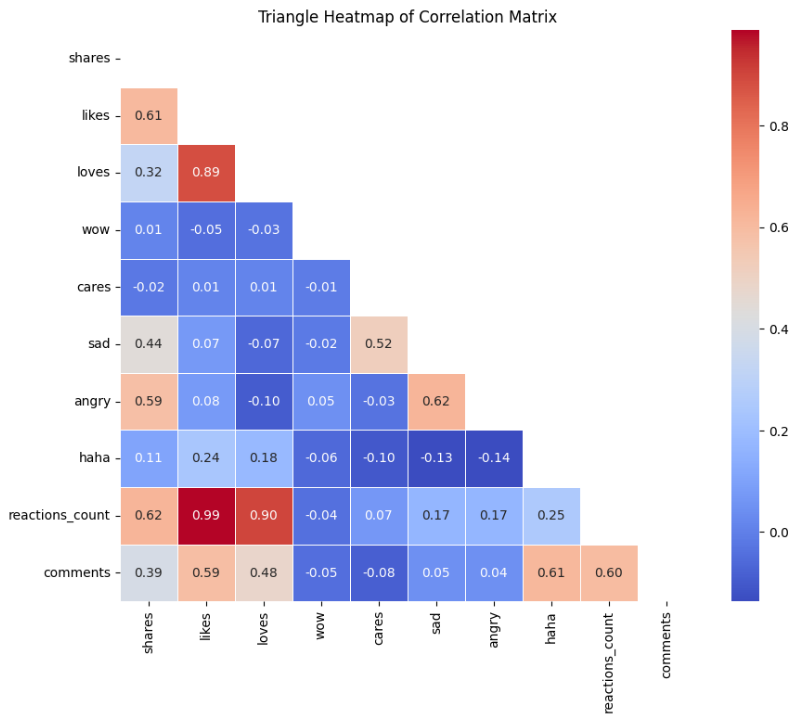
Related to research question 2, another analysis is done to answer the prevalence of negative comments.

- *How does the prevalence of negative comments received by politicians correlate with user engagement metrics, and are there any gender-specific differences in these correlations?*

First, comments are subjected to the "text blob" library, which filters out negative remarks based on sentiment polarity. The average number of comments per post and the correlation between user engagement measures are computed for both male and female politicians. Purpose of this to find the frequency of negative sentiment in user interactions.

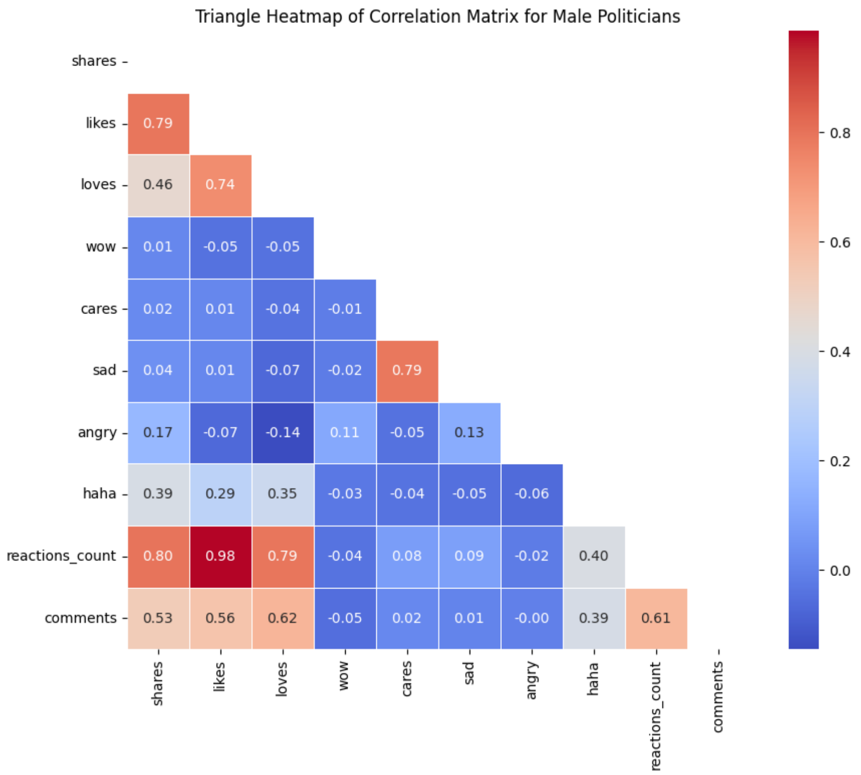
Detailed analyses of female and male politicians' engagement metrics are given below. According to the results, there are minor gender-specific differences in the correlation between negative comments received by politicians and user engagement metrics. A similar pattern can be seen for both genders, which is a weak positive correlation. As these correlations are insignificant, other factors may be beyond negative comments likely to affect user participation in political posts.

Figure 10a: Correlation Matrix for Female Politicians



Female Politicians: The correlation matrix (see Figure 10a) shows moderate to strong positive correlations between various engagement metrics (such as the number of shares, likes, loves, and reactions) and comments. In particular, we observe strong positive correlations between comments and likes (0.591), number of responses (0.597), and shares (0.394). Comment sentiment also shows a weak positive correlation with shares (0.439) and number of reactions (0.170). This indicates that posts with more positive comments tend to receive higher engagement. Interestingly, there is a weak negative correlation between comments and "wow" reactions (-0.048), meaning that posts that receive more "wow" reactions may receive fewer comments.

Figure 10b: Correlation Matrix for Male Politicians



Male Politicians: Similarly, for male politicians, we observe moderate to strong positive correlations between engagement metrics and the number of comments at Figure 8b. Shares (0.530), likes (0.563), and the number of reactions (0.608) show strong positive correlations with comments. The sentiment of comments exhibits a very weak positive correlation with the number of likes (0.012) and number of reactions (0.086); This shows that posts with more positive sentiment comments are slightly associated with higher engagement. However, there is a very weak negative correlation between comments and "wow" reactions (-0.052), indicating that posts that receive more "wow" reactions tend to receive fewer comments.

Overall, the findings show that for both male and female politicians, engagement metrics such as shares, likes, and reactions are positively associated with the number of comments their posts receive. Posts with higher engagement metrics tend to attract more comments; this indicates a potential correlation between engagement levels and audience engagement. However, the sentiment of comments appears to have only a minor impact on engagement metrics; sentiment shows weak correlations with some likes and reactions but no significant correlation with other engagement metrics.

6 Discussion

By understanding the gender-specific dynamics of online toxicity among politicians, targeted interventions and policies can be developed to reduce online harassment and foster a more inclusive and respectful online environment. Policymakers can use insights from this research to design strategies to combat online toxicity and protect the mental health of politicians, particularly female politicians who may face higher levels of toxicity.

6.1. Interpretation of the Findings

According to the thesis results, the following essential findings or patterns have been observed between male and female politicians in online discourse and audience engagement:

Sentiment Distribution: Male politicians received higher sentiment scores, indicating a more polarized and diverse political scene. Comments directed at female politicians, on the other hand, showed a more balanced representation of positive, negative, and neutral sentiments, suggesting a less polarized discourse environment.

Toxicity Levels: In terms of toxicities, female politicians often faced accusations of betrayal, collusion, and treachery in comments directed at them, reflecting a prevalent narrative of mistrust and skepticism. The comments directed at male politicians also contained negative sentiments, including allegations of criminal behavior, corruption, and scandal, highlighting perceived moral lapses and ethical violations.

Engagement Metrics: Regarding engagement metrics (likes, shares, reactions), female politicians received fewer engagement metrics than male politicians, indicating that women politicians are less likely to interact with the audience and be interested in what they are saying. Female politicians received fewer engagements but often received positive comments, such as praise for leadership and integrity, indicating that strong leadership qualities and ethical conduct are recognized.

Topic Discussions: The main topics discussed in comments addressed to male politicians were often serving the people, the need for change, and national issues, reflecting a focus on public service and society. In contrast, topics discussed in comments directed at female politicians frequently centered around making a difference, women's empowerment, and policy initiatives, highlighting a focus on empowerment and policy advocacy.

Correlation with Negative Comments: There was a positive association between the number of negative comments and sad reactions for both male and female politicians, suggesting a slight increase in the probability of users reacting with "sad" feelings as the number of negative comments rises.

Overall, the findings suggest that online discourse surrounding male and female politicians differs in sentiment distribution, toxicity levels, engagement metrics, and topic discussions. Understanding these differences is crucial for addressing gender-based biases and promoting inclusive and respectful dialogue in the digital public sphere.

6.2. Implications of the Study

The findings of this study hold significant implications for various stakeholders, including policymakers, social media platforms, and society at large. By uncovering gender disparities in online discourse, particularly in the domains of politics and social issues, this research sheds light on the need for targeted interventions to address toxic behavior and promote gender equity in online spaces. Additionally, identifying topics with varying sentiment levels provides insights into the public's attitudes and perceptions, which can inform decision-making processes in both the public and private sectors. Furthermore, understanding the differences in sentiment between male and female politicians across different topics can contribute to more informed political communication strategies and public engagement efforts.

6.3. Ethical Considerations

This study adhered to ethical principles and considerations throughout the research process. The data utilized in this study were publicly available and anonymized to protect the privacy of individual users. Moreover, ethical standards were maintained by aggregating and analyzing data at the group level, focusing on average trends rather than individual behaviors. This approach ensured that the results presented in this thesis do not compromise the anonymity or confidentiality of any specific users. Additionally, ethical considerations were considered when interpreting and reporting findings to avoid reinforcing stereotypes or perpetuating harm.

6.4. Limitations

Data Collection Challenges: Social media platforms are characterized by noise, including irrelevant or spam content, complicating data collection. Preprocessing social media data was time-consuming because it involved cleaning, filtering, and normalizing the text, making the analysis challenging.

Challenges in Detecting Sentiment Analysis: Determining the sentiment polarity of text accurately, especially in nuanced or ambiguous contexts, presented a challenge. Identifying sarcasm in the text is difficult as it often involves language that appears positive but conveys a negative sentiment. Analyzing sentiment in multilingual social media data posed challenges due to language-specific nuances and variations.

Challenges in Topic Modeling: Memory limitations have arisen when applying topic modeling techniques to large datasets, leading to errors or crashes. I worked with smaller sample sizes to address memory limitations for some analyses.

7. Conclusion

Understanding and addressing toxic behavior on social media, particularly in discussions involving male and female politicians, is imperative. This study investigates distinctions in online discourse between genders using advanced machine learning and natural language processing techniques.

The evolution of social media has transformed how politicians engage with the public, yet it has also facilitated the proliferation of toxic rhetoric. This study, through an in-depth analysis of sentiment, toxicity, and engagement in comments directed at politicians on Facebook, uncovers significant disparities in how male and female politicians are treated online. These findings not only reveal the urgent need for interventions to mitigate online gender-based violence but also highlight the gravity of the issue.

Key insights from the research include distinct patterns in emotional expression, toxicity levels, engagement metrics, and the topics discussed by male and female politicians. Understanding these nuances is essential for fostering inclusive dialogue and promoting respectful interactions in the digital realm.

Future investigations could delve deeper into how online toxicity influences political polarization and shapes public opinion. The potential of leveraging technological advancements, such as enhanced hate speech detection algorithms, presents promising avenues for addressing online toxicity. This optimistic outlook underscores the importance of collaboration among various stakeholders in developing robust strategies to effectively tackle this pervasive issue.

.

References

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., & Wood, E. (2021, August). Hate speech in political discourse: A case study of UK MPs on Twitter. In Proceedings of the 32nd ACM conference on hypertext and social media (pp. 5–16).

Ali, M. F., Irfan, R., & Lashari, T. A. (2023). Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: a study on governmental preventive measures. PeerJ Computer Science, 9, e1220.

Alkomah, F., Ma, X., & Salati, S. (2022). A literature review of textual hate speech detection methods and datasets. Information, 13(6), 273.

Alkomah, F., Salati, S., & Ma, X. (2022). A new hate speech detection system based on textual and psychological features. International Journal of Advanced Computer Science and Applications, 13(8). https://doi.org/10.14569/ijacsa.2022.01308100

Alshamrani, S., Abuhamad, M., Abusnaina, A., & Mohaisen, D. (2020, October). Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube. In CIKM (Workshops).

Ames, S., & Havens, L. (2022). Exploring National Library of Scotland datasets with Jupyter Notebooks. IFLA journal, 48(1), 50-56.

Anjum, & Katarya, R. (2023). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. International Journal of Information Security, 1-32.

Arango, A., Pérez, J., & Poblete, B. (2019, July). Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval (pp. 45-54).

Awal, R., Cao, R., Lee, R., & Mitrovic, S. (2021). Angrybert: joint learning target and emotion for hate speech detection. https://doi.org/10.48550/arxiv.2103.11800

Awal, R., Lee, R., Tanwar, E., Garg, T., & Chakraborty, T. (2023). Model-agnostic meta-learning for multilingual hate speech detection. https://doi.org/10.48550/arxiv.2303.02513

Bai, Y., & Wang, J. (2015, November). News classifications with labeled LDA. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) (Vol. 1, pp. 75-83). IEEE.

Bacco, L., Russo, F., Ambrosio, L., D'Antoni, F., Vollero, L., Vadalà, G., ... & Denaro, V. (2022). Natural language processing in low back pain and spine diseases: A systematic review. Frontiers in Surgery, 9, 957085.

Boukabous, M., & Azizi, M. (2021). A comparative study of deep learning based language representation learning models. Indonesian Journal of Electrical Engineering and Computer Science, 22(2), 1032. https://doi.org/10.11591/ijeecs.v22.i2.pp1032-1040

Chhablani, G. (2021). Nlrg at semeval-2021 task 5: toxic spans detection leveraging bert-based token classification and span prediction techniques. https://doi.org/10.48550/arxiv.2102.12254

Cheng, J., Tsoh, J. Y., Guan, A., Luu, M., Nguyen, I. V., Tan, R., ... & Burke, N. J. (2022). Engaging Asian American communities during the COVID-19 era tainted with anti-Asian hate and distrust. American Journal of Public Health, 112(S9), S864-S868.

Chiril, P., Pamungkas, E., Benamara, F., Moriceau, V., & Patti, V. (2021). Emotionally informed hate speech detection: a multi-target perspective. Cognitive Computation, 14(1), 322-352. https://doi.org/10.1007/s12559-021-09862-5

Costa-jussà, M. R., & Moreno, A. (2021). Wiser: Weight-sharing for multilingual transformer. https://doi.org/10.48550/arxiv.2104.11338

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology, 7, 886498.

Esposito, E., & Breeze, R. (2022). Gender and politics in a digitalised world: Investigating online hostility against UK female MPs. Discourse & Society, 33(3), 303-323.

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., ... & Al-qaness, M. A. (2021). Social media toxicity classification using deep learning: real-world application UK brexit. Electronics, 10(11), 1332.

Fichman, P., & McClelland, M. W. (2021). The impact of gender and political affiliation on trolling. First Monday.

Helma, C., & Kazius, J. (2006). Artificial intelligence and data mining for toxicity prediction. Current Computer-Aided Drug Design, 2(2), 123-133.

Isaksen, V., & Gambäck, B. (2020, November). Using transfer-based language models to detect hateful and offensive language online. In Proceedings of the fourth workshop on online abuse and harms (pp. 16-27).

Jongeling, R., Sarkar, P., Datta, S., & Serebrenik, A. (2017). On negative results when using sentiment analysis tools for software engineering research. Empirical Software Engineering, 22, 2543-2584.

Karhade, A. V., Lavoie-Gagne, O., Agaronnik, N., Ghaednia, H., Collins, A. K., Shin, D., & Schwab, J. H. (2022). Natural language processing for prediction of readmission in posterior lumbar fusion patients: which free-text notes have the most utility?. The Spine Journal, 22(2), 272-277.

Koc-Michalska, K., & Lilleker, D. G. (2019). Political communities on Facebook across 28 European countries. Questions de communication, (36), 245-265.

Kužel, R., Godársky, I., Kohn, B., & Holly, M. (2022). How Women Politicians of Fiji are Treated on Facebook.

Liboiron, M., Tironi, M., & Calvillo, N. (2018). Toxic politics: Acting in a permanently polluted world. Social studies of science, 48(3), 331-349.

Hatebase. (2020). Retrieved March 01, 2024, from https://www.hatebase.org/

Hatebase Inc. (2021). Hatebase: A database of hate speech. Retrieved March 01, 2024, from https://www.hatebase.org/

Marjanovic, S., Stańczak, K., & Augenstein, I. (2022). Quantifying gender biases towards politicians on Reddit. PloS one, 17(10), e0274317.

Mertens, A., Pradel, F., Rozyjumayeva, A., & Wäckerle, J. (2019, June). As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians. In Proceedings of the 10th ACM Conference on Web Science (pp. 193-201).

Montero, A. I., Laforgue-Bullido, N., & Abril-Hervás, D. (2022). Hate speech: a systematic review of scientific production and educational considerations. Revista Fuentes, 24(2), 222-233.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. PloS one, 15(8), e0237861.

Nockleby, J. T. (1994). Hate speech in context: The case of verbal threats. Buff. L. Rev., 42, 653.

Oates, S., Gurevich, O., Walker, C., & Di Meco, L. (2019). Running while female: Using AI to track how Twitter commentary disadvantages women in the 2020 US primaries. Available at SSRN 3444200.

Ombui, E., Muchemi, L., & Wagacha, P. (2019, October). Hate speech detection in code-switched text messages. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-6). IEEE.

Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. Online Journal of Communication and Media Technologies, 13(4), e202348. https://doi.org/10.30935/ojcmt/13603

Othman, M., Hassan, H., Moawad, R., & Idrees, A. M. (2018). A linguistic approach for opinionated documents summary. Future Computing and Informatics Journal, 3(2), 152-158.

Reuters Institute. (2023). Digital News Report 2023. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Reichert, E., Qiu, H., & Bayrooti, J. (2020). Reading between the demographic lines: Resolving sources of bias in toxicity classifiers. arXiv preprint arXiv:2006.16402.

Qureshi, K. A., & Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. IEEE Access, 9, 109465-109477.

Rheault, L., Rayment, E., & Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. Research & Politics, 6(1), 2053168018816228.

Salminen, A. (2020). Narratives of National Identity: Developing Nationalism in Multicultural Societies.

Samuel-Azran, T., & Yarchi, M. (2023). The “gender affinity effect” behind female politicians' social media support: facebook civil talk during Israel's 2021 elections. Online Information Review, 47(6), 1168-1189.

Sobieraj, S. (2020). Gender, Digital Toxicity, and Political Voice Online.

Sosea, T., & Caragea, C. (2021, August). eMLM: a new pre-training objective for emotion related tasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 286-293).

Stieglitz, S., & Dang-Xuan, L. (2012, January). Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In 2012 45th Hawaii international conference on system sciences (pp. 3500-3509). IEEE.

Solovev, K., & Pröllochs, N. (2022, April). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In Proceedings of the ACM Web Conference 2022 (pp. 3656-3661).

van der Vegt, I. (2023). Gender Differences in Abuse: The Case of Dutch Politicians on Twitter. arXiv preprint arXiv:2306.10769.

Wagner, A. (2022). Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. Feminist Media Studies, 22(1), 32-47.

Wilhelm, C., & Joeckel, S. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. Sex Roles, 80, 381-392.

Appendix

I. The code of the completed work is available at

https://drive.google.com/file/d/1u-LChm1q_IjwhF_eM71UF5Pamwbyxb8K/view?usp=drive_link

II. Figures

Figure 11: Sentiment Distribution for Comments Directed at Male and Female Politicians

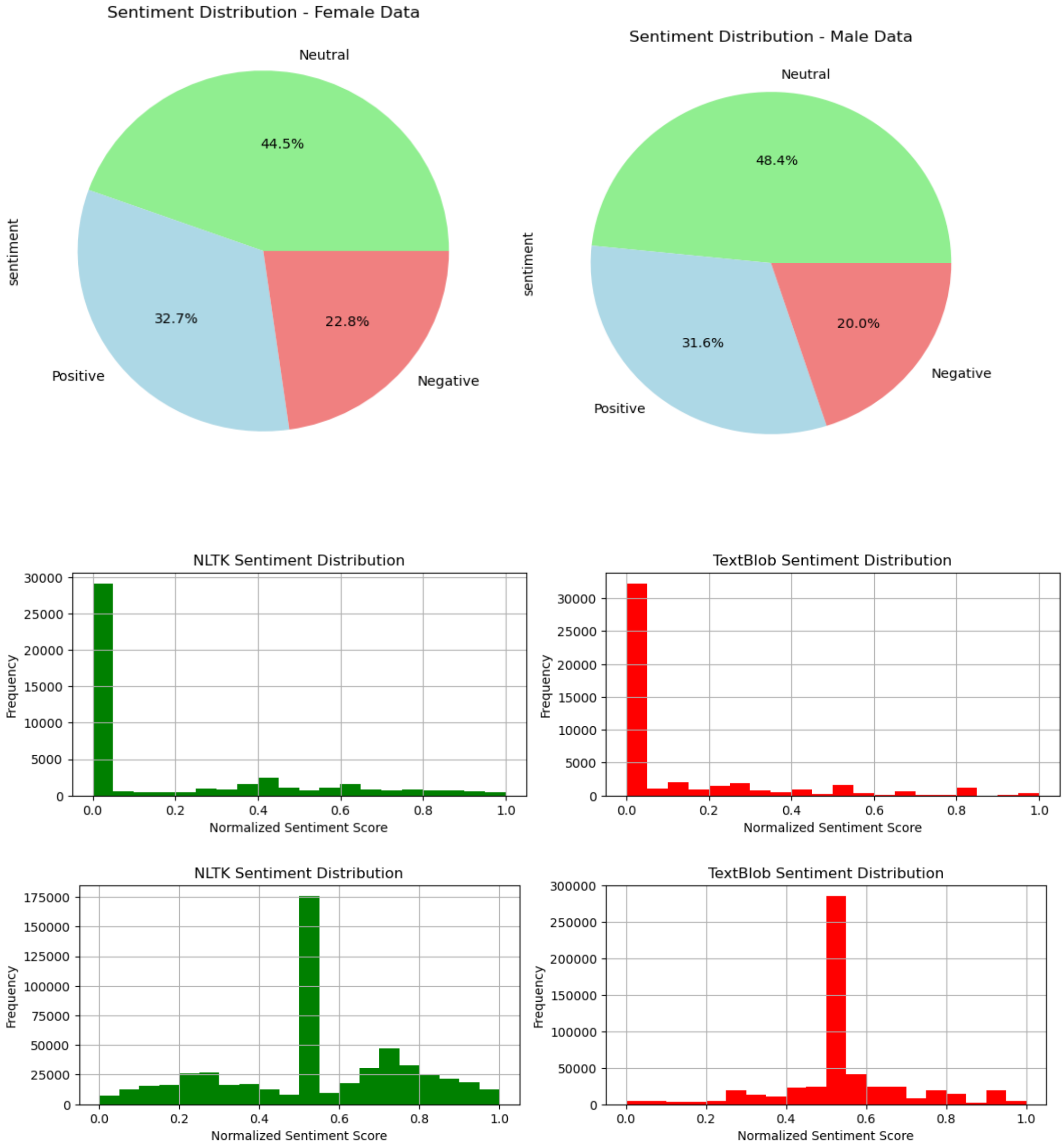
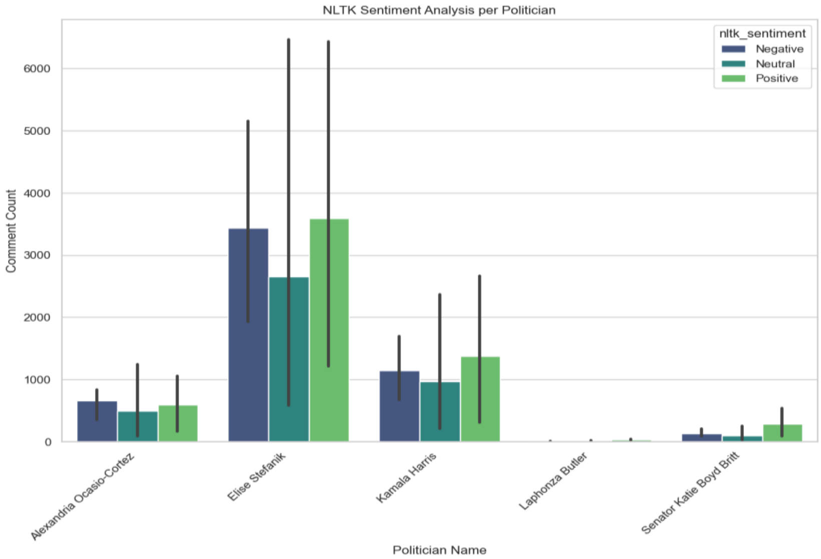


Figure 12: NLTK and TextBlob Sentiment Distribution for Male and Female Politicians' Comments



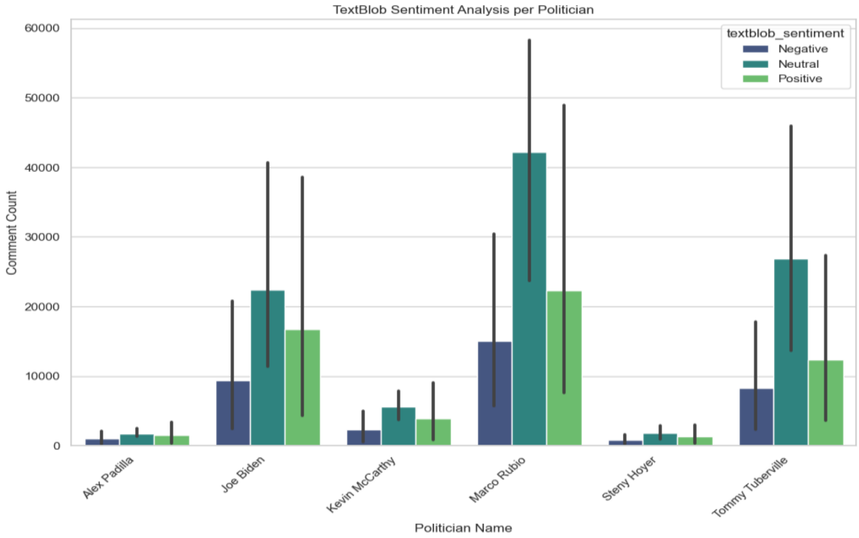


Figure 13. NLTK and Textblob sentiment analysis per politician of the male dataset

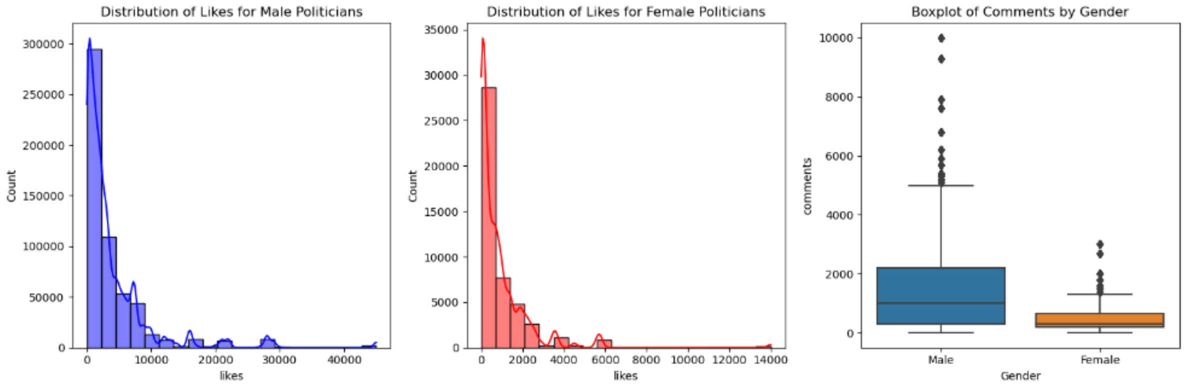
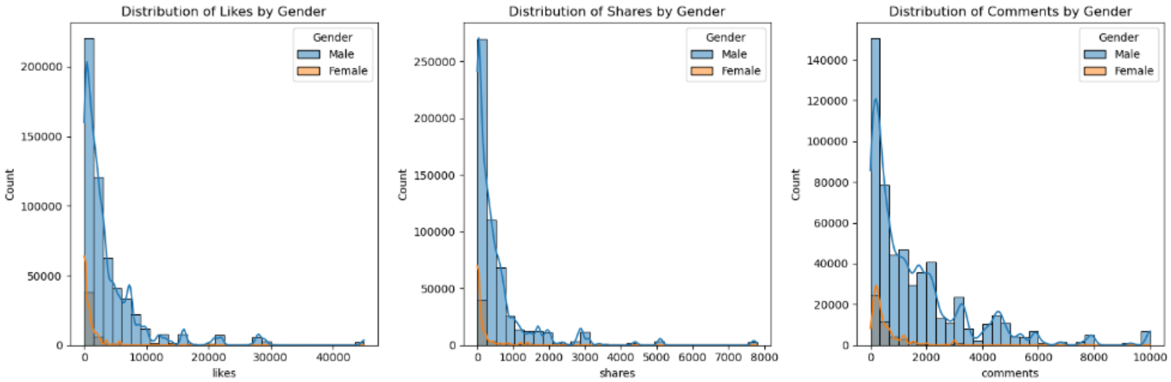


Figure 14. Distribution of emotions by Male and Female Politicians



III. License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Devrim Nesipoglu

1. grant the University of Tartu a free permit (non-exclusive licence) to:

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Comparison of toxicity among female and male active politicians in social media,

supervised by Uku Kangur and Rajesh Sharma,

2. I grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from ***16/05/2024*** until the expiry of the term of copyright,
3. I am aware that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Devrim Nesipoglu

15/05/2024