Utrecht University

Department of Information and Computing Science

Applied Data Science Master's Thesis

“What Has Been Said Cannot Be Taken Back”:

A Toxic Speech Detection Framework for TikTok using Whisper and Perspective API

First Examiner:
Dr. Jing Zeng

Candidate:
Jelle Prins, BSc
Std. Nr. 5875196

Second Examiner:
Dr. Dennis Nguyen

July 7, 2023

Abstract

On social media platforms, such as TikTok, toxic speech is a common problem. With a focus on videos from the 2020 US presidential election, this study suggests a framework for spotting toxic speech in TikTok videos. For the purpose of transcribing and analyzing spoken content in TikTok videos, the framework combines a speech-to-text algorithm and a toxicity detection API.
The findings show that TikTok videos have varying amounts of toxic speech, with the majority of texts scoring low for toxicity. With the help of BERTopic, semantic characteristics extraction, dominant topics like Joe Biden's actions and discussions of race and politics are identified. Sentiment analysis shows different emotional tones across topics. It is also shown that there may be a correlation between some sentiments and higher levels of toxicity by looking at the relationship between toxicity and sentiment. These findings provide insights into the characteristics of toxic speech in TikTok videos. The results contribute to the development of strategies for content moderation and the promotion of healthier online communities. Future research should address limitations and further explore toxic speech on video-based social media platforms.

Keywords: toxic speech, TikTok, social media, toxicity detection, speech-to-text algorithm, sentiment analysis.

Table of Contents

Abstract 2
Table of Contents 3
1. Introduction 4
2. Related work 5
2.1 Toxicity detection in text-based social media 5
2.2 Toxicity detection in video-based social media 5
3. Methodology 6
3.1 Dataset 7
3.2 Dataset processing 7
3.3 Results of the data collection and processing 8
3.4 Analyses 10
3.4.1 Toxicity Analysis 10
3.4.2 Extracting Semantic Characteristics 11
3.4.3 Combining Analyses 12
3.4.4 Political Polarization in the data 12
4. Results 13
4.1 Findings from the toxicity analysis 13
4.2 The findings from the semantic characteristics extraction 14
4.3 Findings of the combined Analyses 16
4.4 Findings from the Analysis of Political Polarization in the data 20
5. Discussion 22
6. Conclusion 24
References 25
Appendix 28
Figures 28

1. Introduction

Since its launch in 2017, TikTok has established itself as a popular social media platform for sharing short videos. TikTok is well-known for its viral dance challenges and short comedy sketches, and it allows users to express themselves or showcase their talents while interacting with their peers in a way that is creative in contrast to traditional text-based social media platforms.
Many people turned to TikTok during the COVID-19 pandemic as a way to communicate with one another while remaining safe in their own homes. However, as the platform grew in popularity, toxic speech also became more common. Like many other social media platforms, TikTok recognizes the negative effects of toxic speech and has policies in place to protect its users. According to TikTok, hate speech and hateful behavior "attack, threaten, dehumanize, or degrade an individual or group based on their characteristics. These include things like race, ethnicity, national origin, religion, caste, sexual orientation, gender, and gender identity, as well as things like serious illness, disabilities, and immigration status (TikTok, 2022). The growing frequency of toxic speech and its potential to negatively affect communities require the creation of robust and trustworthy systems that can identify it and protect TikTok users from its effects.

Although toxicity detection in online posts is not a new area of study and several social media platforms already have algorithms in place that address toxic posts, the special characteristics of the short videos posted on TikTok require the development of a specialized framework that can successfully detect and analyze toxicity in speech in this context. TikTok videos are primarily made up of spoken words and audio elements like sound effects or music snippets. Finding toxicity in TikTok videos is challenging due to these elements.

In this thesis, a new framework for identifying toxic speech in spoken text within TikTok videos is proposed. By combining a highly effective speech-to-text algorithm with recent developments in automatic speech recognition and natural language processing, the framework aims to accurately transcribe and analyze spoken content in TikTok videos and find instances that contain toxic speech.
The framework will allow the application of well-known text-based toxicity detection techniques to the transcribed speech data by turning the audio elements of TikTok videos into textual representations. This approach can help us better understand the toxic usage of speech that appears in TikTok videos and allow us to develop specific measures that will contribute to a TikTok community that is both healthier and more diverse.

In the context of researching toxic speech in TikTok videos, it is also important to take into account specific events that have had a major impact on society as a whole. For example, social media platforms like TikTok saw a spike in political engagement and discussion during the 2020 U.S. presidential elections (Medina Serrano et al., 2020). It is impossible to ignore TikTok's influence on political narratives and the growth of political speech. This thesis focuses on TikTok videos from the 2020 U.S. presidential election with the goal of understanding how toxic speech appears on TikTok, especially in the political domain.

This thesis advances wider studies on toxicity detection and content moderation on online video-based platforms by addressing the difficulties of toxicity detection in speech on TikTok. The findings of this thesis can be used to create efficient pipelines and tools that can be customized for similar (short) video-based social media platforms, helping to decrease the negative effects of toxic speech on those platforms.

In the sections that follow, we will review relevant literature on toxicity detection on social media and on video-based social media platforms, describe the method used to create the suggested framework, go over the experimental results, and then offer suggestions on how to implement and improve the suggested framework for TikTok and other similar platforms.

2. Related work

2.1 Toxicity detection in text-based social media

The issue of identifying hate speech has given rise to multiple approaches in recent years. Different deep learning and machine learning algorithms have been used to implement these solutions.

On the basis of deep learning classifiers and word embeddings, D'Sa et al. (2020) proposed a novel method for automatically detecting toxic speech in social media. Both binary and multi-class classifications were studied by the authors. Toxic and non-toxic speech were used for the binary classification, and hate speech, offensive speech, and neither were taken into account for the multi-class classification. FastText and BERT embeddings, both feature-based methods, were used as input to CNN and Bi-LSTM classifiers in the method. These approaches' classification performance was evaluated against a pre-trained, fine-tuned BERT model. It was found that for the provided Twitter dataset, the fine-tuned BERT model outperformed the feature-based approaches.

Singh et al. (2022) draw attention to the growing issue of toxicity, vulgarity, and cyberbullying in online platforms, particularly social media. In addition to using individual models like Naive Bayes, Logistic Regression, SVM, and SVM, the authors also used ensemble models like Random Forest and XGBoost. Deep learning models like LSTM and GRU were also used because the dataset was so large (about 48,000 tweets). The models' F1 scores, which show the superior performance of deep learning models in this scenario, ranged from 0.84 for Random Forest to 0.92 for GRU and 0.91 for LSTM. Based on factors like age, gender, ethnicity, and religion, the experiment's findings demonstrate its potential for accurately identifying cyberbullying.

2.2 Toxicity detection in video-based social media

While there has been a lot of research on identifying hate speech, it has largely focused on textual data, such as comments, posts, blogs, tweets, etc. Research must be done to learn how to recognize hate speech in videos because they can also be used to spread hate speech. Videos have been categorized using a variety of techniques, including machine learning and manual annotation. The use of existing toxicity detection tools for classification, however, has received relatively little research.

Hernandez Urbano Jr. et al. (2021) suggested using BERT to identify toxicity in transcriptions from Tagalog TikTok videos. 1000 TikTok videos were manually transcribed, and each video was annotated as toxic or not. The data was used to train the Filipino BERT model as well as a number of other machine learning models. The optimal model for the dataset was selected using a weighted average of Micro and Macro F1 scores. The experimental outcomes showed that the Bernoulli Naive Bayes model outperformed the Filipino BERT model in the detection of hate speech, obtaining a significantly higher Weighted average F1 score of 74% in contrast to BERT's score of 62%.

By transcribing the spoken words into text before providing them as input to machine learning models, Wu and Bhandary (2020) implemented a method to detect hate speech in YouTube videos. Four different machine learning models were trained by the authors to categorize videos into three groups: normal, racist, and sexist. The best working model was chosen using evaluation metrics such as accuracy, precision score, recall, and F1 score. The Random Forest Classifier, which scored 96% on accuracy, was found to deliver the best outcomes.

Vasconcellos et al. (2023) proposed a framework to identify toxicity and polarization in political debate on TikTok. For the purpose of creating a less noisy dataset, the framework uses a robust audio cleaning pipeline. Despite the difficulties of spoken text, the authors were able to extract coherent and important topics from TikTok using tools like a web crawler, an audio segmentation tool, a speech-to-text algorithm, and topic modeling. Qualitative analysis revealed that topics pertaining to religion and social classes had higher concentrations of toxic videos and polarization.

3. Methodology

This thesis proposes a framework for identifying toxic videos on TikTok based on a combination of textual features from metadata and spoken content in videos. The framework is based on the following steps:

1. Extract metadata and videos from TikTok.
2. Separate the parts of TikTok videos that contain human speech using speech detection and audio segmentation tools.
3. Convert the extracted speech into text format using a speech-to-text algorithm.
4. Compute toxicity scores for each video using a deep learning-based toxicity detection API.

Fig. 1 provides a comprehensive visual representation of the process used to create this framework. The appendix contains a larger version of this figure.

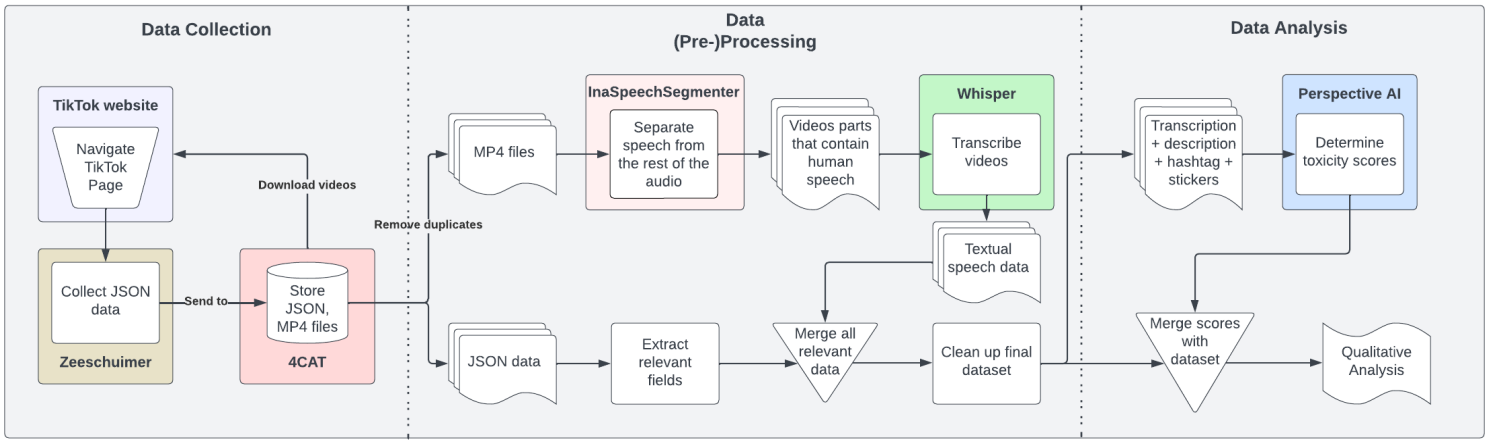

Fig. 1 Methodology for the construction of the dataset and subsequent analysis

3.1 Dataset

Due to the lack of a readily accessible dataset, the information for toxic and non-toxic TikTok videos was manually gathered. Since only researchers based in the United States have access to TikTok's API, the data had to be scraped from the company's website. Based on the most widely used hashtags during the election on Twitter as discovered by Chen et al. (2022) and Ferrara et al. (2020), six TikTok hashtag pages for the 2020 US presidential elections were selected. The hashtags were associated with both the Democratic Party and the Republican Party, to ensure a diverse dataset.

Zeeschuimer, a tool that tracks internet traffic while browsing, was used to create the dataset (Peeters, 2023). Zeeschuimer gathered the metadata for every video associated with each hashtag by manually scrolling through the hashtag pages. This metadata was subsequently sent to an instance of 4CAT running on a local machine. 4CAT is a research tool that enables the analysis and processing of data from various online social platforms (Peeters & Hagen, 2022). The videos associated with each metadata entry were downloaded and saved on a local drive using 4CAT's web interface.

3.2 Dataset processing

After collecting the MP4 videos and their corresponding JSON-formatted metadata for each hashtag, a Python script was used. This script combined the data and organized the MP4 files into a single folder, using a unique video ID found in the metadata entries. A second Python script was then executed to remove duplicate videos and metadata entries, ensuring that each video only appeared once.

To identify any unneeded audio or noise in the MP4 videos and remove it from the data, inaSpeechSegmenter was used to process the videos. Convolutional neural networks are used in this audio segmentation toolkit, which can distinguish between speech, music, and noise as well as the gender and voice activity of the speaker (Doukhan et al., 2018). The sections of each video that included sound, speech, or music were labeled in this step. The pydub library, an audio manipulation library with a straightforward and high-level interface, was used to extract only the audio portions of the video using the timestamps for each of these parts. The resulting audio files were saved in MP3 format for further processing.

The development of this framework made use of Whisper, a speech-to-text library created by OpenAI (Radford et al., 2022). The MP3 audio files containing human speech were converted into text using Whisper. Whisper's ability to identify the language being spoken in an MP3 file is one of its benefits. This made it possible to further clean the dataset so that only English transcriptions were generated. The word error rate (WER) for Whisper's performance on the English Fleurs dataset was only 4.2%, which was the third-best result (OpenAI, 2022). The dataset was prepared for the last stage of data processing after the transcriptions were combined with the corresponding metadata from the videos.

3.3 Results of the data collection and processing

During the manual collection process, a sizable number of TikTok videos linked to particular hashtags associated with the 2020 US presidential election were retrieved. Table 1 lists all of the hashtags chosen for data collection along with how many videos were manually retrieved for each one.

Table 1: Hashtags selected for data collection

Hashtag	**Political affiliation**	**Number of videos retrieved**
trump2020	Republican	903
biden2020	Democrat	1039
maga	Republican	935
joebiden	Democrat	862
trump	Republican	1044
democrats	Democrat	1130
	Total	**5913**

These figures highlight the significant effort put forth to compile a broad and comprehensive dataset for the analysis of toxic speech during the 2020 US presidential election. A thorough investigation into the frequency and characteristics of toxic speech across party affiliations was made possible by the inclusion of multiple hashtags associated with both the Democratic Party and the Republican Party. This guaranteed a diverse representation of political conversations. The large number of videos that were found for each hashtag shows the effort put into gathering a variety of viewpoints and opinions from TikTok users during this significant period in politics. For understanding the patterns of toxic speech during the 2020 US presidential election, this dataset is an extremely helpful tool.

Several filtering techniques were used to further refine the dataset during the processing stage. One such technique for identifying and separating videos that only contained music or noise was the inaSpeechSegmenter tool. The videos were divided into sections based on sound, speech, and music so that segments irrelevant to the analysis could be found and removed. With the help of this filtering procedure, a dataset with more specific video and audio clips was produced. An example of the audio waveform representation for the output generated by inaSpeechSegmenter is shown in Fig. 2. The figure shows the audio line, which represents the strength of the audio signal, while various colors highlight the music, male, and female speech segments. The time labels are used to distinguish the various audio waveform segments, even though they are not depicted in the figure. Fig. 2 is included to help with understanding the segmentation procedure and shows how the speech segmenter correctly recognizes and distinguishes between various audio components in the videos.

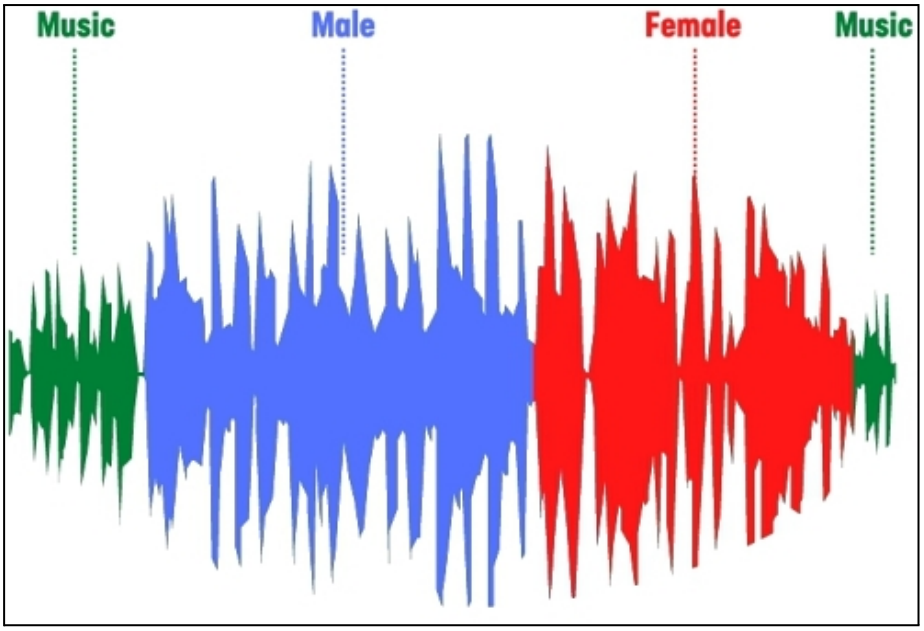

Figure 2 An example of the output generated by inaSpeechSegmenter. The framework provides segments' time spans and their labels. Taken from Vasconcellos et al. (2023).

Overall, the careful steps taken during the dataset processing stage, including the modification and organizing of the TikTok videos, allowed for a thorough review of toxic speech during the 2020 US presidential election. By manually gathering a large number of videos linked to relevant hashtags, a diverse dataset reflecting multiple points of view and ideas has been created. The dataset was filtered using a number of techniques, including the use of inaSpeechSegmenter to detect and classify audio segments. The result was a narrower selection of video clips with relevant audio content. The dataset was ultimately significantly shrunk down to a final set of 1052 TikTok videos that satisfied specific criteria and were ready for analysis. In order to capture the context surrounding the 2020 US presidential election, these criteria included choosing videos from a two-year time span, covering one year before and one year after the election. Furthermore, only videos with human speech were included, ensuring that there were no videos with only music or noise. Additionally, language filtering was used to ensure that the spoken language in the videos was English. Following these strict criteria, the final dataset produced a precise and relevant set of TikTok videos that made it possible to carry out a thorough investigation of toxic speech among those affiliated with the Democratic and Republican parties during the election.

3.4 Analyses

3.4.1 Toxicity Analysis

The first step involved using the Perspective API to evaluate the text's toxicity, which included a combination of the video description, speech transcription, stickers used in the video, and hashtags. Perspective uses machine learning models to detect toxic comments. Based on their predictions of how a text will influence a conversation, the models assign a text a score. Toxic language can be filtered out with the aid of this score (Jigsaw, 2023). On a scale from 0 to 1, the probability scores for toxicity are given, with 1 denoting that the text is extremely likely to be toxic and 0 denoting that it is not likely to be toxic. In addition to toxicity, a number of other characteristics are evaluated on a comparable scale; a summary of these characteristics and their meanings is given in Table 2. The dataset was prepared for qualitative analysis of the results after each score was added to its corresponding video.

Table 2 Attribute names measured by Perspective and their descriptions[1]

Attribute name	**Description**
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Severe Toxicity	A user is likely to leave a discussion or give up sharing their point of view if someone makes a comment that is extremely hateful, aggressive, disrespectful, or other similar things. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Identity Attack	Negative or hateful comments targeting someone because of their identity.
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
Profanity	Swear words, curse words, or other obscene or profane language.
Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.

Jigsaw's Perspective API served as the main tool for this study's analysis of text toxicity and hate speech for a variety of reasons. First of all, Perspective AI provides a free tier that gives users access to the necessary functionality without having to pay any additional charges. The rate at which one can send API requests is constrained, but given the small number of documents that remained after preprocessing (N = 1052), this constraint is not relevant for this thesis. For our research needs, it is therefore a good option. Secondly, Perspective API is trusted and used by well-known businesses such

[1] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

as Disqus, The Financial Times, The New York Times, The Wall Street Journal, and Reddit. This proves the industry's wide acceptance of it and its dependability. Finally, Jigsaw's Perspective API provides seamless Python integration via the Google API library. Once configured, it offers simple instructions for integrating its features into our research workflow. Jigsaw's Perspective API emerged as the best option for our study after taking these factors into account because it offers free access, has a trustworthy user base, and has simple integration with the Python-based research environment.

Alternative solutions for text toxicity analysis that were taken into account during the evaluation process included OpenAI's Content Moderation API, Sift's Content Moderation API, Two Hat's Community Sift API, and the Google Cloud Natural Language API. However, these alternatives required paid services, which is not regarded as a good research practice.

Additionally, Ye et al. (2023) contend that, when used on unclean data, Perspective API provides the most reliable scoring outcomes when compared to cutting-edge language models like BERT and RoBERTa. Due to its availability, robustness against uncleaned data, accessibility, and simple integration into the research workflow of this thesis, Jigsaw's Perspective API was chosen in the end.

3.4.2 Extracting Semantic Characteristics

To extract semantic characteristics, a variety of natural language processing (NLP) methods were applied to the video transcriptions. These techniques included topic modeling and sentiment analysis.

Topic Modeling

Due to its adaptability, stability across domains, and support for multilingual analysis, BERTopic was chosen for topic modeling (Egger & Yu, 2022). BERTopic automatically calculates the number of topics and does not require the original data to be preprocessed. Additionally, it has built-in search capabilities and hierarchical topic reduction. But it could also produce an excessive number of topics, produce outliers, and assign documents to a single topic. Metrics for objective evaluation are also lacking. Despite these drawbacks, BERTopic was effective at identifying the semantic features of the toxic videos and offering insights into their underlying themes and topics.

Sentiment Analysis

The emotional undertone of the speech in the TikTok videos was identified through sentiment analysis, which could be used to understand the potential impact the speech might have on the viewer. VADER (Valence Aware Dictionary and Sentiment Reasoner) was the tool of choice for sentiment analysis in this study. VADER is commonly used to analyze social media texts because it is capable of handling the details and informal expressions frequently found online (Hutto & Gilbert, 2014). VADER was used in an effort to accurately represent the sentiment found in the toxic videos, taking into account both the context and the features that define the content. Polarity analysis can help further this research's goal of identifying the characteristics of toxic speech in TikTok videos by providing insight into the personal and emotional impact of the material being studied.

3.4.3 Combining Analyses

The goal of the combined analyses sections in this study is to give readers a thorough understanding of toxic speech in relation to various topics, the connections between topics, the interaction between sentiment and toxicity, and the linguistic patterns connected to toxic behavior. By examining toxicity by topic, topic co-occurrence, sentiment-toxicity relationship, and word/phrase analysis, we can learn a lot about the frequency, dynamics, and characteristics of toxic speech, which will help us develop targeted interventions, content categorization strategies, sentiment-aware moderation tools, and precise toxicity detection models. Collectively, these analyses support the development of safer online communities and the encouragement of positive online discourse.

Toxicity by Topic

To understand the frequency and pattern of toxic speech across different topics or themes in the data, toxicity must be analyzed by topic. We can learn which topics are more likely to provoke toxic behavior or produce harmful content by looking at the distribution of toxicity scores for each topic. This analysis enables us to pinpoint potential toxic hotspots and rank interventions or preventative measures accordingly. In addition, breaking down the toxicity by topic makes it easier to comprehend the effects of particular topics on online discourse and facilitates the creation of targeted moderation strategies.

Sentiment-Toxicity Relationship

To understand how emotional tone interacts with the presence of toxic speech, the relationship between sentiment and toxicity must be explored. We can determine which sentiments are more likely to coexist with toxicity or serve as a mitigating factor by looking at the correlation between sentiment scores and toxicity scores. This analysis sheds important light on how the likelihood of toxic behaviors is influenced by positive or negative sentiments, shedding light on the complex dynamics of online conversations. The creation of sentiment-aware moderation tools, sentiment-based content filtering, and proactive intervention strategies can benefit from an understanding of the relationship between sentiment and toxicity.

3.4.4 Political Polarization in the data

To investigate how toxic speech affects the data's bipartisan nature, several visualizations were created. This was accomplished by contrasting the toxicity ratings for the two parties, which were represented by the hashtags associated with each video. In the same way that the hashtags "joebiden," "democrats," and "biden2020" were combined into a single tag, "Democrats," so too were the hashtags "trump," "trump2020," and "maga," which were combined into a single tag, "Republicans." Following the creation of these labels, the distribution of toxicity sub-attributes between the two parties was also compared and inspected. Finally, the relationship between TikTok's social media metrics and toxicity was visualized in order to examine how they relate to each party. This thesis aims to shed some light on the connection between political polarization and the frequency of toxic speech during the 2020 US presidential election through the use of the analyses presented here.

4. Results

The TikTok videos contain toxic speech, and the qualitative analysis of the dataset provided insightful information about its nature and characteristics. A good grasp of the qualitative aspects of toxic speech during the 2020 US presidential election was reached through an extensive study of the content, patterns, and context-specific details.

4.1 Findings from the toxicity analysis

The distribution of toxicity scores sheds light on the degree of toxicity seen in the texts in the analysis. If a video transcription receives a score of at least 0.7, it is deemed toxic. The frequency of various toxicity scores given to the texts is shown in the histogram plot in Fig. 3. The y-axis shows the frequency or count of texts falling within each toxicity score range, and the x-axis represents the toxicity score, which ranges from 0 to 1. The histogram shows that, with a peak around 0.1, the majority of texts have relatively low toxicity scores. This shows that the majority of the texts that were examined have a low level of toxicity. The frequency of texts does, however, gradually increase as the toxicity score approaches the range of 0.3 to 0.4. The kernel density estimation (KDE) line also offers more details about the distribution's form and smoothness. A gradual transition from lower to higher toxicity scores is indicated by the KDE line, which suggests a relatively smooth distribution. These results show that the analyzed texts have varying levels of toxicity, with the majority being non-toxic or toxic to a very low degree.

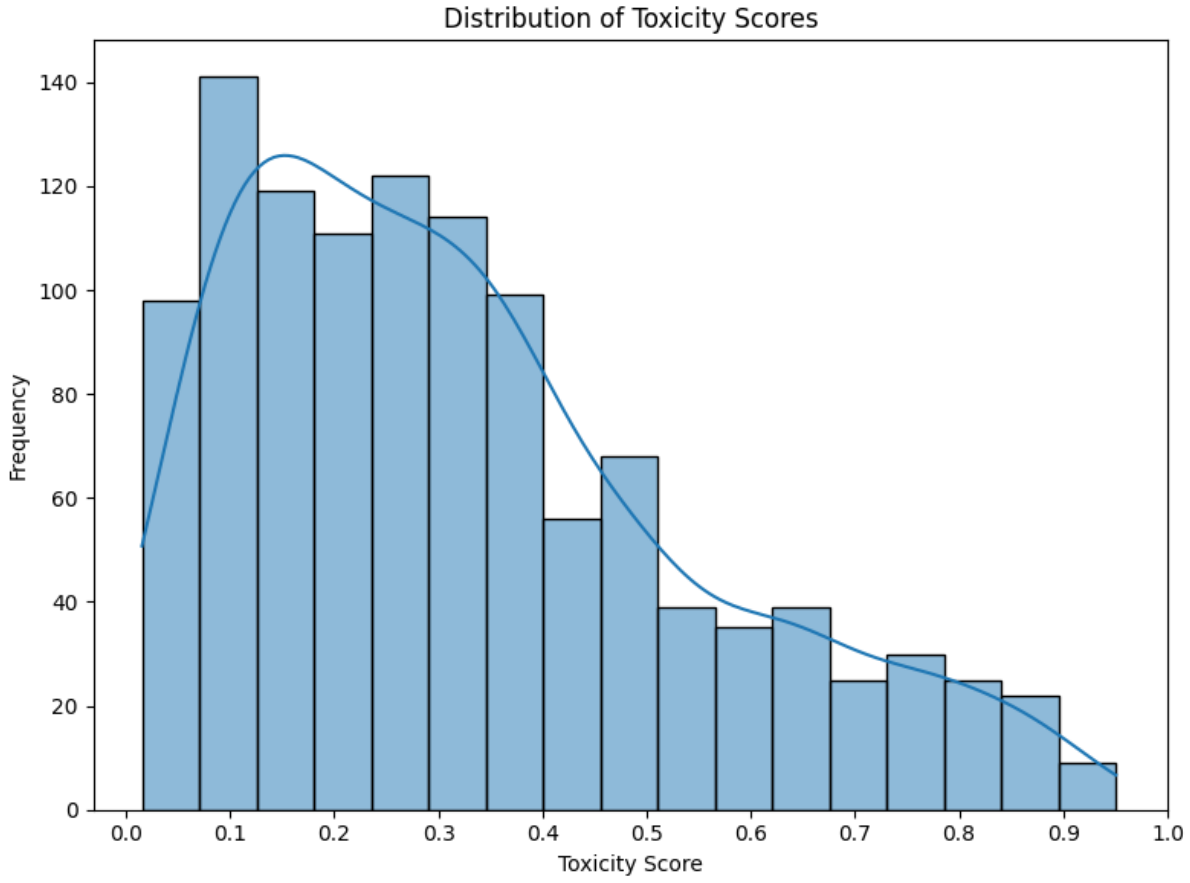

Figure 3 Histogram of the distribution of toxicity scores found by the Perspective API

The results of the toxicity analysis will be used in the sections that follow to provide additional insight on the characteristics of toxic speech.

4.2 The findings from the semantic characteristics extraction

Topic modeling was done using BERTopic using the following model parameters:

```
topic_model = BERTopic(
    nr_topics="auto",
    embedding_model="all-MiniLM-L6-v2", # best model for general purposes
    verbose=True
)
```

With the aid of these parameters, BERTopic was able to automatically determine the ideal number of topics while using a robust language model to produce document embeddings. The topic modeling procedure is described in great detail when verbose mode is enabled. These decisions helped to produce topics from the input documents that were meaningful and understandable. Following the execution of BERTopic, a total of 13 topics were discovered; Table 3 lists the number of documents associated with each topic. The topic names were manually selected after carefully reading the "representative documents" that BERTopic discovered. The first notable outcome is the topic evaluation of -1 given to 478 documents. This shows that the BERTopic model did not assign the topics to any particular topic. This might be because the document is an outlier, the documents do not have enough context or information to be assigned, or the model's confidence in the topic assignment may be low given the clustering and similarity measures on which BERTopic is based. Topics with IDs 0 and 1 are also noteworthy results because they both contain more than 100 documents, suggesting that they may be the transcriptions' dominant themes.

Table 3 Topics discovered by BERTopic and manually selected topic names

Topic ID	Document Count	Topic Name
-1	478	None
0	168	Joe Biden's Behavior
1	113	Discussions on Race and Politics
2	77	Post-Election Uncertainty and Political Divide
3	73	Pro-Trump Social Media Posts and Reactions
4	72	Criticism and Defense of Trump and Biden Supporters
5	28	Fictional and Real Personal Experiences
6	27	Miscellaneous TikTok videos
7	27	Criticism of Trump's Handling of the Corona Virus
8	23	Companies that Donated to Political Candidates
9	20	Discussions Around Wearing Face Masks
10	17	Commentary on Trump's Family
11	17	Trump and Biden Related Products and Merchandise
12	12	Climate Change and Environmental Policies

The VADER sentiment analysis tool was used to analyze the sentiment scores for the transcriptions of TikTok videos. Vader, a model for sentiment analysis of social media text, is frequently used because it is good at capturing the intricate details of sentiment in short and informal texts (Hutto & Gilbert, 2014). A deeper understanding of the topic's sentiment polarity is made possible by the integration of topic modeling and sentiment analysis, which provides insightful information about the emotional characteristics of transcriptions linked to particular topics. This combined strategy improves the overall analysis of the data by enabling a more complex understanding of the emotional tone displayed in each topic. The boxplot in Fig. 4 shows the sentiment scores assigned to each topic determined by the BERTopic model. The sentiment score, which ranges from 1 (indicating a positive sentiment) to -1 (indicating a negative sentiment), is represented on the vertical axis. The plot's boxes each represent the interquartile range, which shows the range of sentiment scores for each topic. The average sentiment score for each topic is indicated by the central line within each box.

Interesting insights about the sentiment scores across various topics are revealed by the boxplot analysis. There is a noticeable difference in the emotional tone among the topics, as shown by the distribution of the boxes representing the interquartile range. The highest average sentiment score for

Topic 6 stands out, indicating a generally positive sentiment surrounding the information within this topic. However, there are some negative outliers present as well. Topic 1 has a lower average sentiment score, which indicates a more emotional tone that is comparatively negative. The average sentiment scores for topics 10 and 11 are relatively high, indicating a largely positive sentiment associated with these topics. The greater variance in sentiment scores for Topic 12 suggests that attitudes toward environmental issues and climate change are subject to a wider range of emotions.

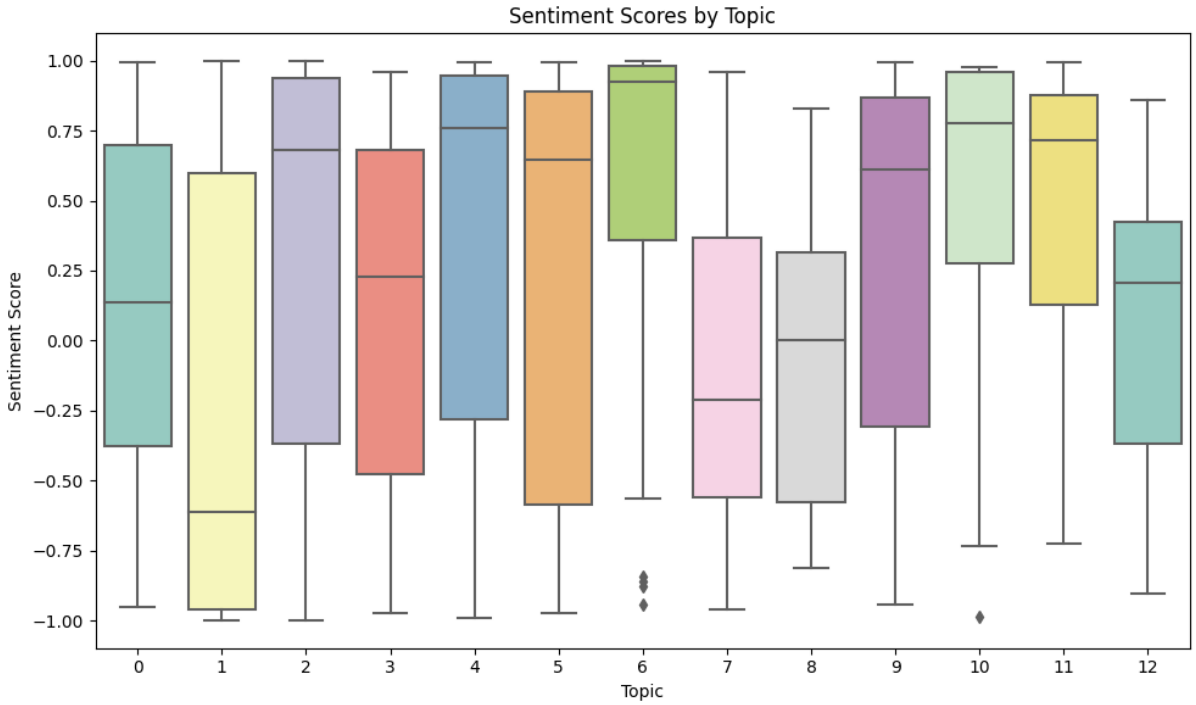

Figure 4 A box plot depicting the sentiment scores associated with each topic identified by the BERTopic model

4.3 Findings of the combined Analyses

Toxicity by Topic

The distribution of toxicity scores for various topics is displayed in Fig. 4's boxplots. The topics relating to "Race and Politics" and "Fictional and Real Personal experiences" contain a higher average of toxic speech, as can be seen from the higher median values in the box plots. Although the topic "Joe Biden's Behavior" has a low median toxicity score, it is noteworthy that quite a few outliers above the top whisker contain toxic speech and can be seen to be above the established toxicity threshold. It is also evident from the box plot that some topics do not at all contain toxic speech; for example, "Climate Change and Environmental Policies," "Trump and Biden Related Products and Merchandise," and "Commentary on Trump's Family" all remain below the toxicity threshold line. For additional analysis into figuring out the characteristics of toxic speech, it is interesting to look at the categories that have outliers or whiskers above the toxicity threshold.

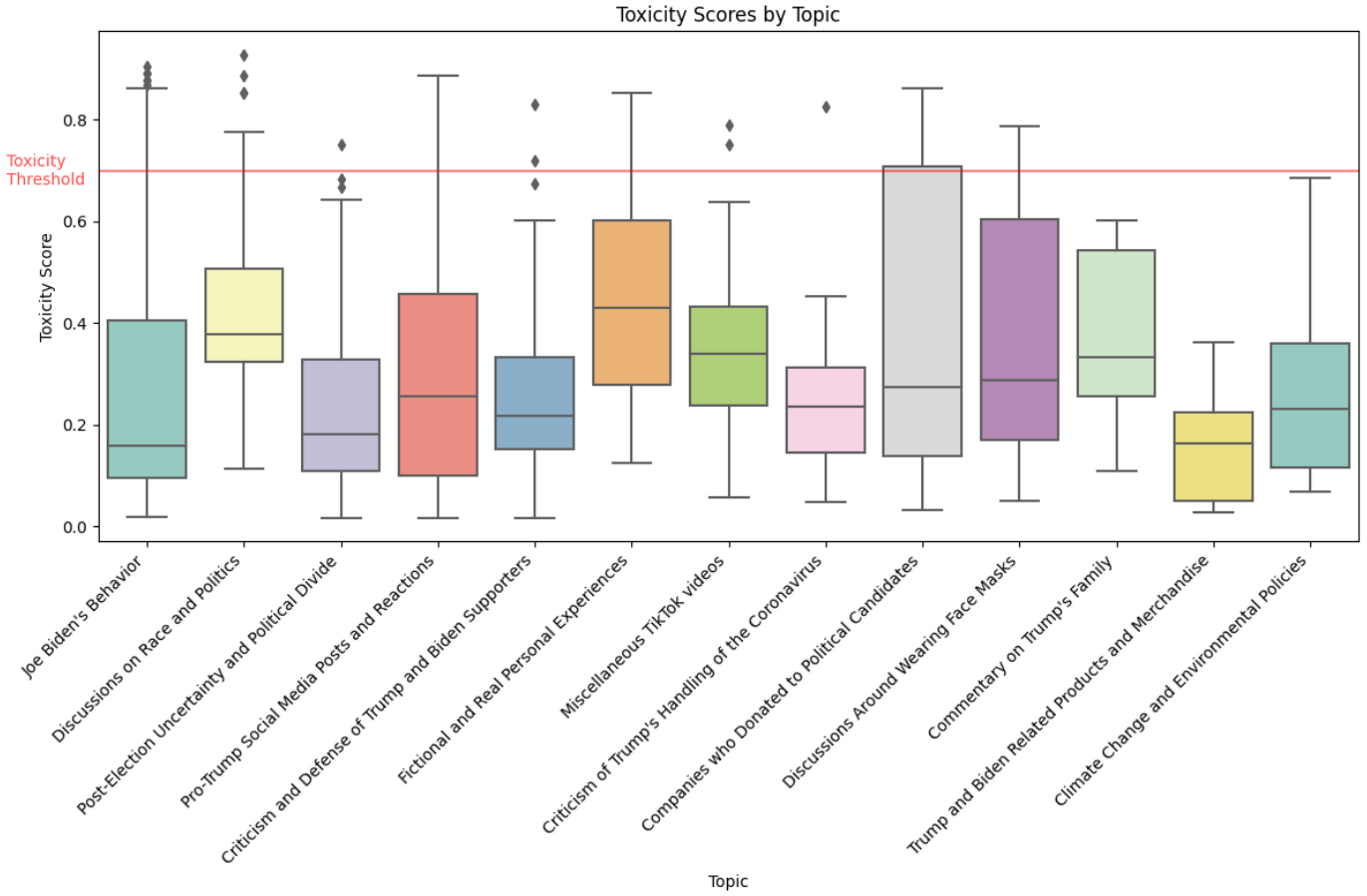

Figure 5 A box plot of toxicity scores for each topic with a red line indicating the toxicity threshold of 0.7

When you examine the results for the toxic videos in more detail, the distribution drastically alters. The box plots in Fig. 6 make it clear that the last three topics, as previously mentioned, do not contain any toxic speech. On the other hand, the topics with a median toxicity score > 0.8 are 1, 2, 4, and 9, indicating that there are a significant number of videos with toxic speech.

A word cloud created for topic 1 is displayed in Fig. 7, and it helps to further illustrate how toxic speech is used. A word cloud is a graphic representation of text data where each word's size reflects how frequently it appears in the text. Larger font sizes are used to visually indicate words that are used more frequently in the text.

When attempting to identify toxic speech within a topic, a word cloud can provide a quick overview and visualization of the most common or significant words used in the documents connected with that topic. By making a word cloud of the documents that are related to the topic, we can identify the key words or phrases that increase the topic's overall toxicity. The words "bitch," "fucking," "fuck," and "trumpsuck" demonstrate the clear use of curse words in this word cloud and are likely responsible for the high toxicity score.

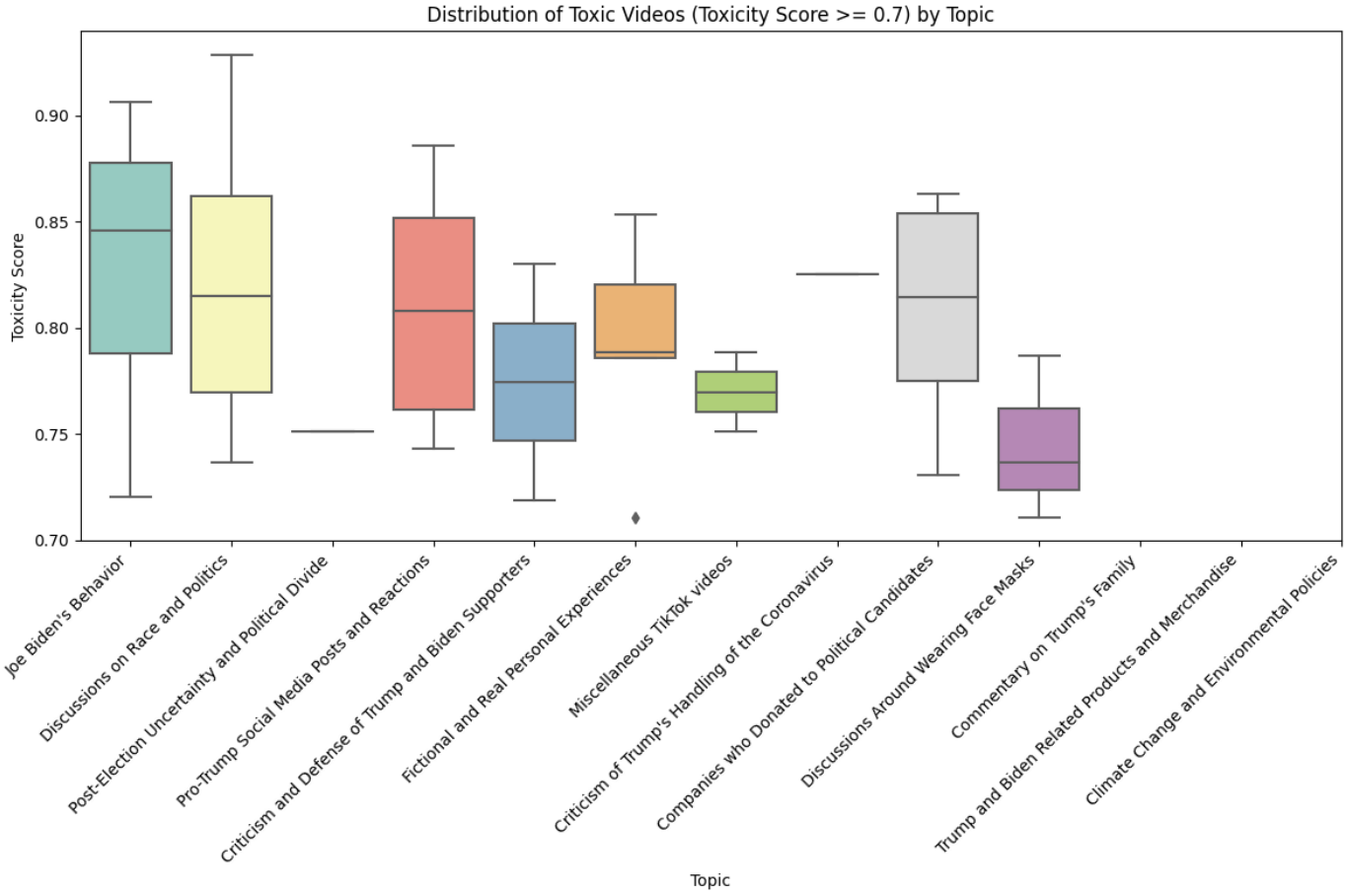

Figure 6 A box plot with the distribution of toxicity scores for toxic videos

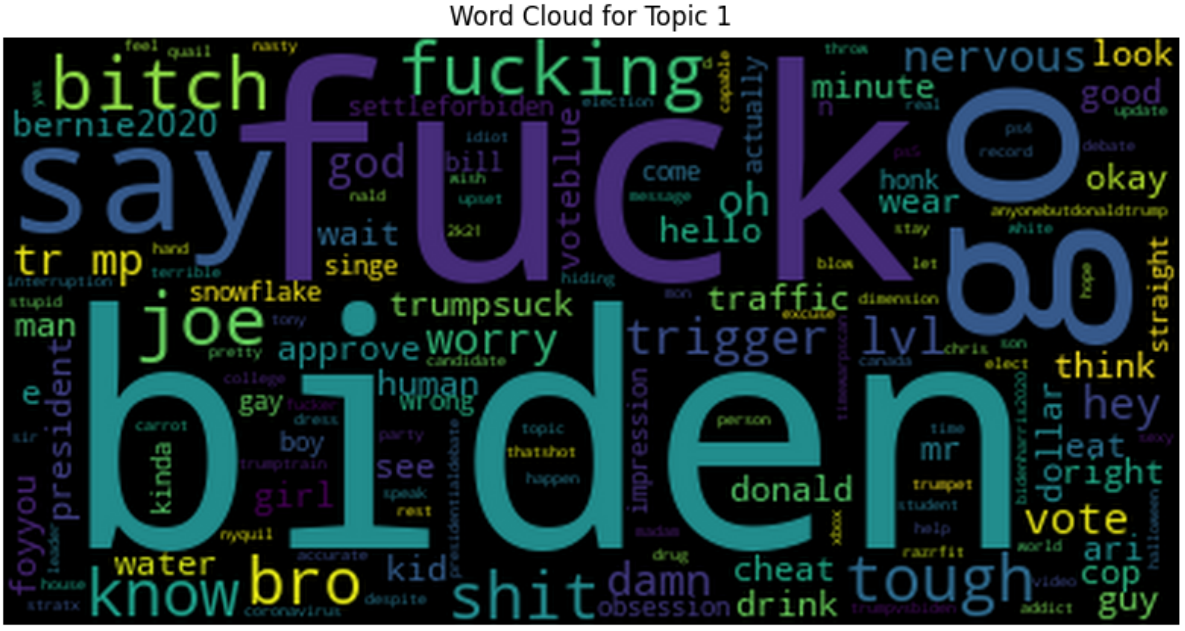

Figure 7 A word cloud generated for toxic documents (toxicity > 0.7) related to Topic 1: "Joe Biden's Behavior"

Sentiment-Toxicity Relationship

In order to determine whether sentiment could be another indicator for toxicity, a scatter plot is made to visually examine the correlation between toxicity scores and sentiment scores. Figure 8 uses the distribution of the two variables on the plot to demonstrate this relationship. Four distinct groups can be seen when the scatter plot is analyzed. The plot's upper left corner displays high toxicity and low sentiment scores, indicating instances of highly toxic language expressed with unfavorable sentiment. In contrast, the low sentiment and low toxicity scores in the bottom left corner of the plot indicate language that is neutral or non-toxic but still has a negative sentiment. As we move to the bottom middle of the plot, we notice a cluster of documents with neutral sentiment (0.0) and low toxicity scores (< 0.4). These documents could indicate neutral or objective speech. The bottom right of the plot, which shows low toxicity scores but high sentiment scores, shows instances of positive sentiment and the absence of toxic language, indicating favorable speech. While the majority of the data points fall into these groups, the plot's top right corner shows a few outliers. These outliers are documents with high sentiment scores as well as high toxicity scores, indicating speech that is both highly positive in sentiment and at the same time contains toxic speech. Overall, by highlighting distinct patterns and potential connections between these two measures, the scatter plot reveals that there is probably no significant connection between sentiment and toxicity.

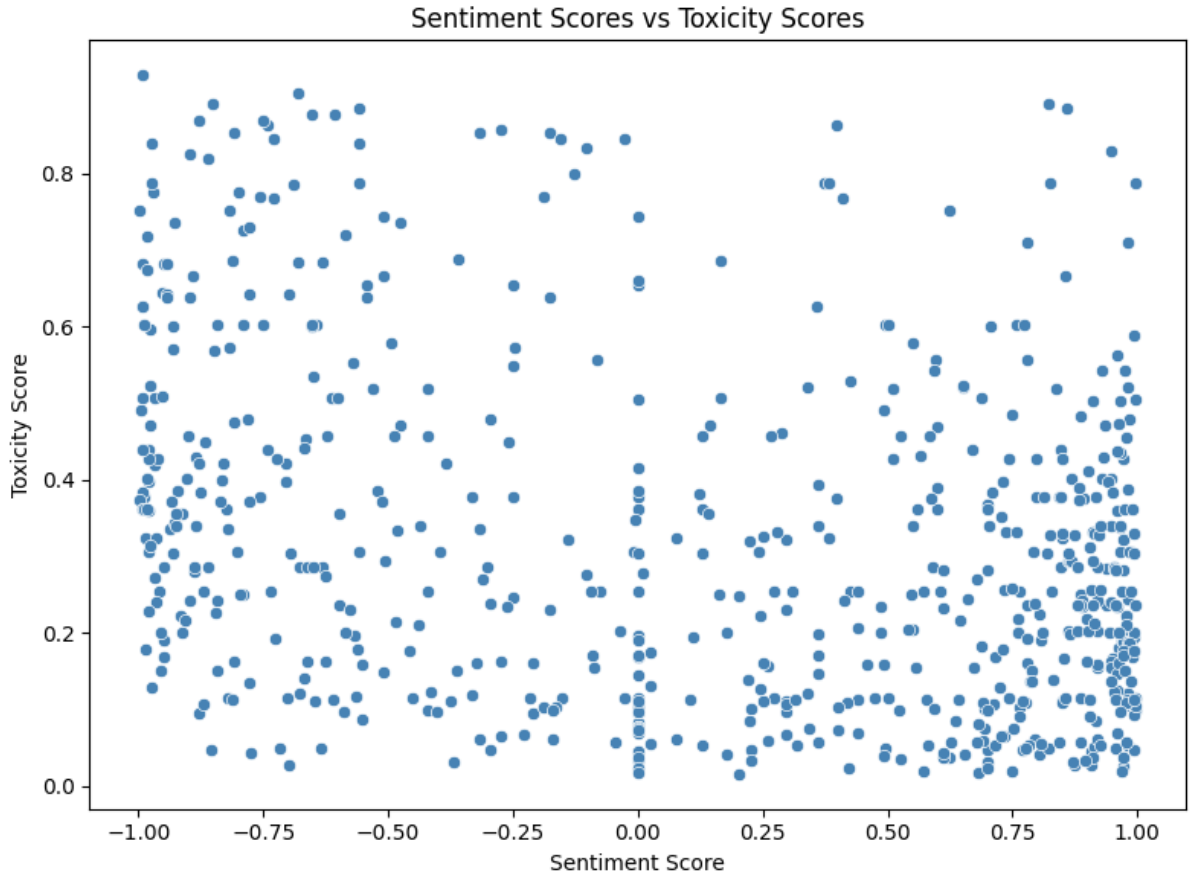

Figure 8 A scatterplot of Sentiment Scores vs Toxicity Scores

4.4 Findings from the Analysis of Political Polarization in the data

The distribution plots of the toxicity score for videos associated with both Democratic and Republican party affiliations are shown in Fig. 4. In order to gain insight into the effect of political polarization on the frequency of toxic speech, the analysis was focused on finding whether there was a difference in the amount of toxic speech used in these videos during the 2020 US presidential election.

Fig. 9 illustrates this with the red distribution representing videos connected to the Republican party and the blue distribution connecting them to the Democratic party. The threshold value (0.7) for a video to be classified as toxic is indicated by the dashed black line. The distribution of toxicity scores for both party affiliations is shown in the histograms, allowing for a visual comparison of the frequency of toxic speech. For videos related to the Democratic party, the average toxicity score is 0.31, while for videos related to the Republican party, it is 0.34. Both distributions have comparable shapes, with the Republican distribution having a slightly longer right tail. In the left tail, the Democratic distribution is a little higher. Overall, there is a leftward shift in both distributions, which indicates a lower frequency of high toxicity scores. Additionally, only a small portion of all videos exceed the 0.7 threshold and fall within the toxic area of the plot. This analysis points out the relatively lower frequency of toxic speech in both groups while also pointing out small differences in the distribution of toxicity scores between the Democratic and Republican party videos.

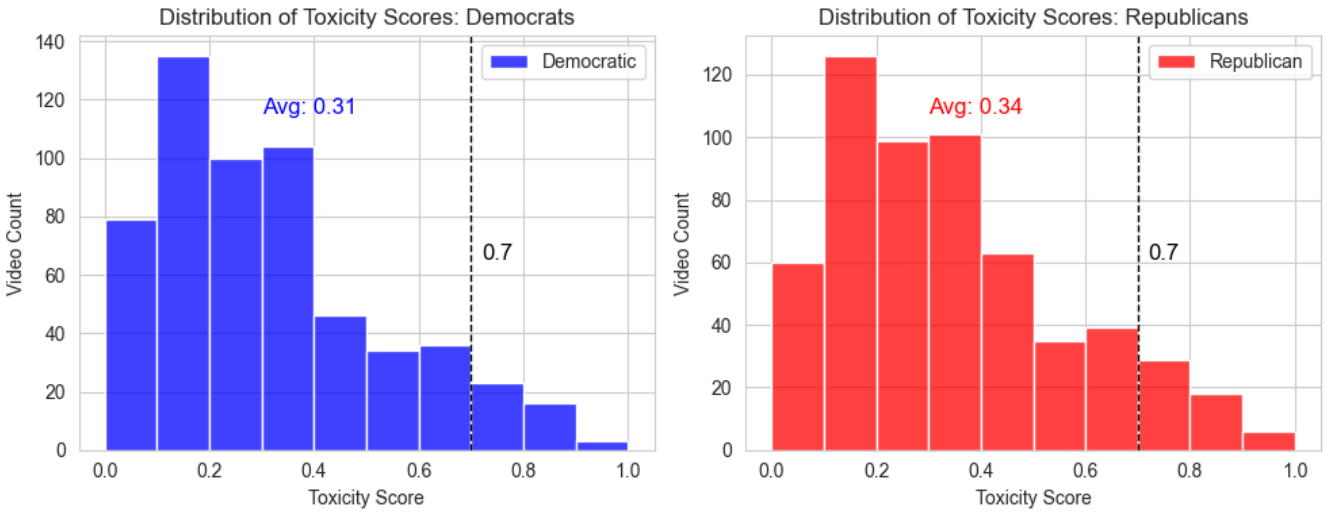

Figure 9 Distribution plots of the toxicity score for videos related to both party affiliations. The black line displays the threshold value for a video to be labeled as toxic.

The distribution of the most common toxicity subcategories within the Democratic and Republican parties is shown in Fig. 10. This figure supports the earlier analysis (Fig. 4) by focusing specifically on the toxic videos (N = 95) and examining the frequency of the subcategories. After reviewing the toxic videos, it was found that profanity accounted for 81.0% of the toxic videos within the Democratic Party, making it the most common subcategory of toxicity. Insults were the second-most common subcategory, accounting for 16.7% of the toxic videos, while identity attacks made up only 2.4%. Similarly, within the Republican Party, profanity accounted for 75.5% of the toxic videos, making it the most common subcategory of toxicity. With 24.5% of the toxic videos falling into this subcategory, insults were the second most common category. These findings highlight the prevalence of profanity in both groups by showing how frequently both party affiliations use explicit language to

express toxic behavior. The distribution discovered further emphasizes the significance of considering specific subcategories in order to develop a deeper comprehension of toxic speech in political conversations. Overall, Fig. 10 highlights the different types of toxic speech that are typical within each political affiliation by comparing the breakdown of toxicity subcategories between the Democratic and Republican parties.

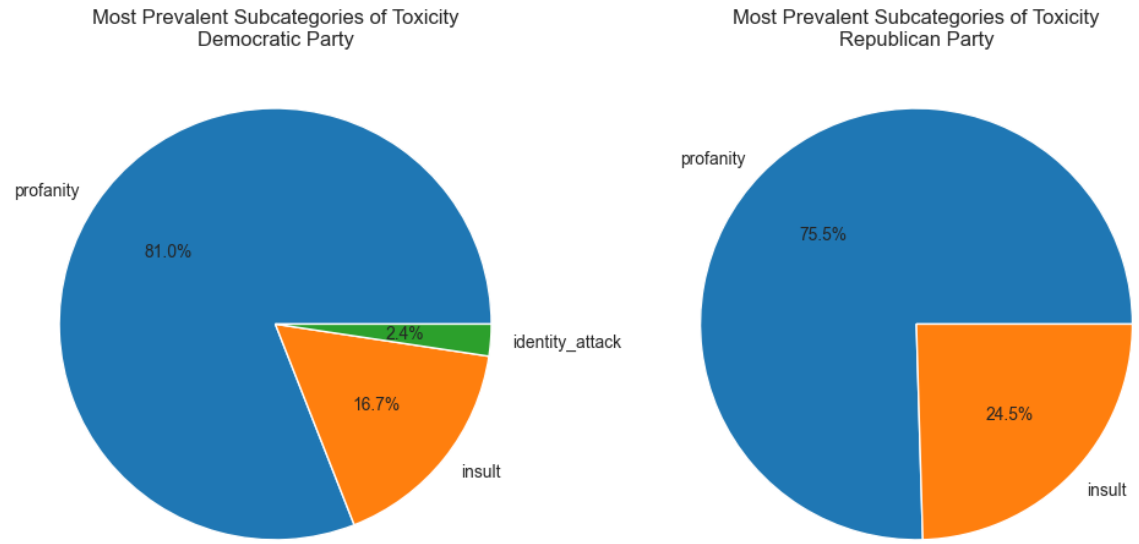

Figure 10 Distribution of the most common subcategories of toxicity in the Democratic and Republican Parties

Plotting the TikTok social media metrics against the toxicity score and dividing them by party affiliation allowed for a final comparison. The end result of this visualization is shown in Figure 11. Because the metrics for some videos range in the hundreds of thousands while only being a few thousand for others, the y-axis is in a logarithmic scale. Smoother visualization is guaranteed when using the logarithmic scale. The four metrics include Diggs, which are similar to likes on other social media platforms, share count, which represents how many times the video was shared, comment count, which represents how many comments viewers have left on the video, and play count, which represents how many times the video has been viewed. Both the democratic party (in blue) and the republican party (in red) have roughly similar distributions for all four metrics. The majority of videos can be seen to fall into the not toxic category for all metrics (toxicity 0.7), as has previously been demonstrated in other visualizations. The spread or variability of the number of Diggs appears to be slightly higher for the democratic party than the republican party, which is one notable distinction between the Diggs plot and the other plots. The equal spread around the center of the graph suggests that there is not a clear correlation between each metric and the toxicity score. Therefore, the metrics are not reliable characteristics of videos with toxic speech.

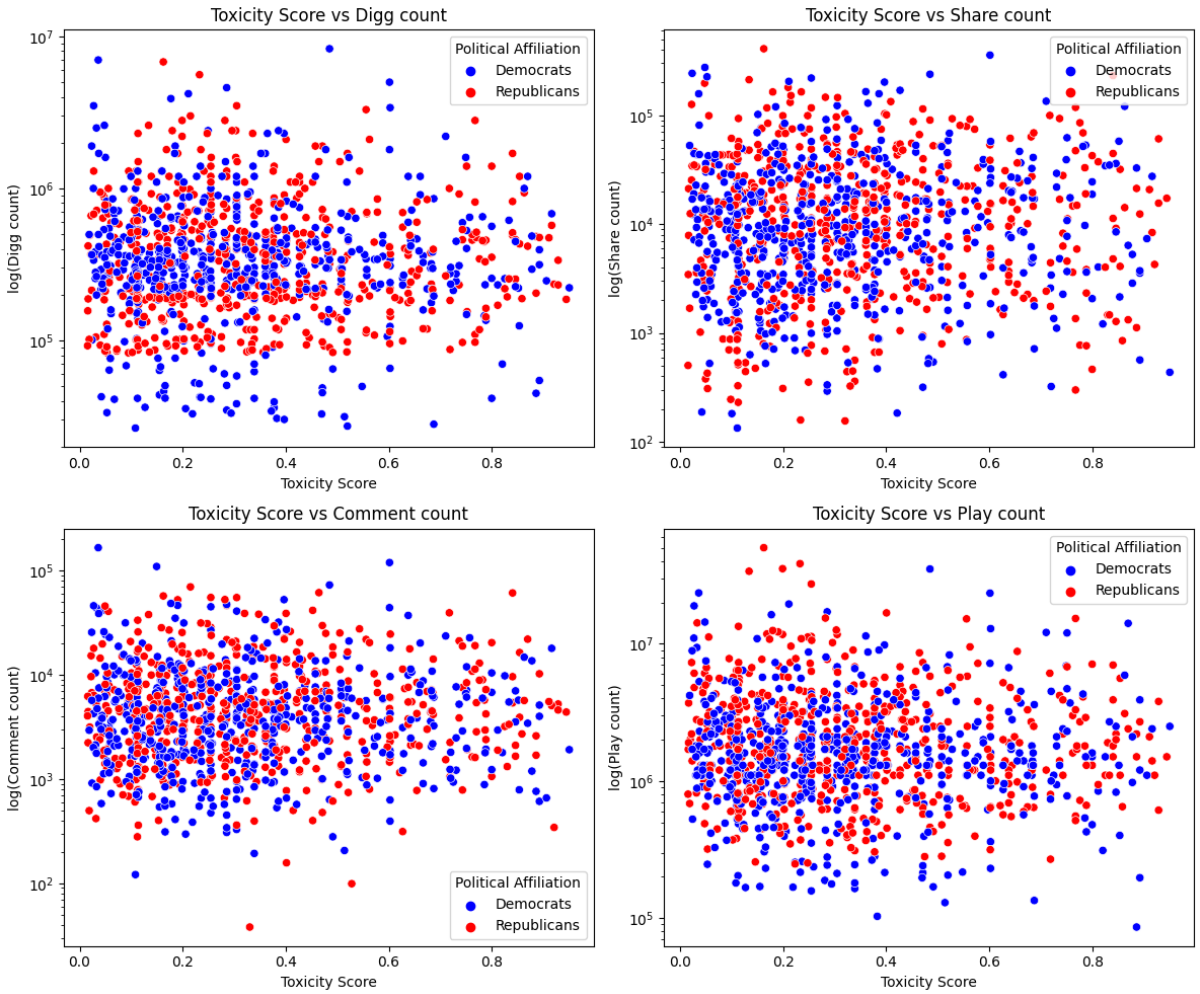

Figure 11 Scatter plots of TikTok's Social Media Metrics Diggs (Likes), Shares, Comments and Number of Plays vs Toxicity scores

5. Discussion

The findings of this study shed important light on the characteristics of toxic speech in TikTok videos. The results of the toxicity analysis showed a range of toxicity scores, which represented the amount of toxicity found in the examined texts. Most of the examined texts showed relatively low toxicity scores, with a peak around 0.1, indicating that their levels of toxicity are generally low. However, as the toxicity score approached the range of 0.3 to 0.4, the number of texts gradually increased. The kernel density estimation (KDE) line shows a change from lower to higher levels of toxicity at this point. In general, the texts under analysis showed varying degrees of toxicity, with the majority falling into the non-toxic or low-toxicity range.

The results of the semantic characteristic extraction, which used BERTopic for topic modeling, improved our comprehension of the content found in the TikTok videos. 13 topics, each associated with a different number of documents, were automatically determined by the BERTopic model. Notably, topics 0 and 1—each of which contained more than 100 documents—emerged as the dominant themes. Joe Biden's behavior, as well as issues of race and politics, were discussed in these topics. The BERTopic analysis also uncovered a group of documents (Topic ID of -1) that were not assigned to any particular topic. This may be due to a number of things, such as the documents being outliers or not having enough context for topic assignment. Using VADER, topic modeling and

sentiment analysis together allowed for a deeper comprehension of the emotional characteristics related to each topic. Topic 6 displayed a generally positive sentiment, while topics 1, 10, and 11 displayed a more negative sentiment, according to the sentiment scores. The sentiment analysis emphasized the variety of attitudes and emotions that were expressed in the transcriptions that were connected to particular topics.

The boxplots of toxicity scores for each topic revealed variations in the presence of toxic speech when analyzing the relationship between toxicity and topics. The higher median toxicity values indicate that topics relating to race and politics, as well as fictional and actual personal experiences, showed a higher average of toxic speech. There were notable outliers with toxic speech above the predetermined toxicity threshold, despite the topic "Joe Biden's Behavior" having a low median toxicity score. Contrarily, the toxicity threshold was not reached for topics like "Climate Change and Environmental Policies," "Trump and Biden Related Products and Merchandise," and "Commentary on Trump's Family," indicating the absence of toxic speech. Additional information about the characteristics of toxic speech within particular categories can be gained by identifying topics with outliers or whiskers above the toxicity threshold.

A significant change in the distribution of the toxic videos was discovered after further investigation. The box plots made it obvious that the final three topics, which had previously indicated the absence of toxic speech, continued to be consistent. The median toxicity score for topics 1, 2, 4, and 9 was higher than 0.8, indicating that there were a significant number of videos with toxic speech within these topics. Word clouds were created to help people understand how toxic speech is used within a topic. The use of words such as "trumpsuck," "bitch," and "fucking" in the word cloud for topic 1, "Joe Biden's Behavior," was an important factor in the topic's high toxicity score.

The analysis also looked at the connection between sentiment and toxicity. To investigate the relationship between the toxicity and sentiment scores, a scatter plot was created. The scatter plot showed four different groups, demonstrating the connection between toxicity and sentiment. The plot's conclusions imply that some sentiment levels might be a sign of higher toxicity levels, while others might correspond to lower toxicity levels. This relationship reveals more about the intricate interactions between sentiment and toxic speech in TikTok videos.

The findings of this study demonstrate the existence of toxic speech in TikTok videos and give a more detailed understanding of its characteristics. The results of the toxicity analysis, semantic characteristic extraction, and sentiment-toxicity relationship analysis provide insight into the distribution of toxicity scores, dominant topics, sentiment, and the connection between sentiment and toxicity. These results contribute to understanding the toxic speech that affects TikTok communities online.

This study has some valuable insights, but there are a few limitations that need to be acknowledged. The focus on TikTok videos specifically related to the 2020 US presidential election limits the findings' generalizability, to start. Depending on the context, the subject, and the user demographics, toxic speech can take on different forms and exhibit different traits. To improve the generalizability of the results, future research should aim to include a wider variety of TikTok videos.

The manual data collection method applied in this study is another drawback. The reliance on using a user account to scroll through TikTok pages introduces potential biases and restrictions through TikTok’s recommendation algorithm. Certain videos might have been missed in the data gathered through manual browsing, which may not have included the full spectrum of TikTok videos. A more extensive and objective dataset for analysis might be available if a web crawler is used. Future research should think about using more organized and automated data collection techniques.

Furthermore, the dataset that was gathered for analysis in this study did not have any manual tags or labels applied. There are uncertainties and restrictions introduced by this reliance on the black box

algorithm of the Perspective API for the toxicity analysis, such as the validation of the toxicity analysis. The lack of a labeled dataset may have an impact on the validity and accuracy of the toxicity assessment. Human annotation or labeling of the dataset would give the toxicity analysis a more solid foundation. For more accurate assessments of toxicity, future research should take into account manual labeling procedures.

Additionally, due to time restrictions, speech-to-text solutions and alternative topic modeling methodologies were not investigated. Although BERTopic and the selected speech-to-text algorithm were successful for this study, other models or algorithms might offer alternative viewpoints and insights. Future studies should investigate additional approaches to improve the analysis of toxic speech in TikTok videos and deepen our understanding of it.

This study advances our knowledge of toxic speech in TikTok videos despite its limitations. The findings shed light on the distribution of toxicity scores, dominant topics, emotional tones, and the relationship between sentiment and toxicity. The framework for toxic speech identification that has been proposed has implications for dealing with the problem and promoting a safer online environment. However, future research should consider addressing the identified limitations to enhance the thoroughness and validity of findings in this area of study.

6. Conclusion

In conclusion, this study offers insightful information about the characteristics of toxic speech in TikTok videos. The results show that the videos under study have varying degrees of toxicity, with the majority falling into the non-toxic or low-toxicity categories. Semantic characteristics extraction using BERTopic for topic modeling improves our understanding of the topics and sentiment within TikTok videos, with topics related to race and politics as well as fictional and personal experiences exhibiting a higher average of toxic speech. The connection between sentiment and toxicity further shows the complex character of toxic speech in TikTok videos.

The limitations of this study must be acknowledged, though. Because of the emphasis on TikTok videos relevant to the 2020 US presidential election, the findings' generalizability is limited. The manual data collection procedure and lack of manual labeling for the dataset introduce potential biases and uncertainties into the analysis. Additionally, due to time restrictions, alternative topic modeling methods and speech-to-text solutions were not investigated.

Despite these drawbacks, this research adds to our understanding of toxic speech in TikTok videos and lays the groundwork for addressing the growing issue of toxic speech on video-based social media platforms. The framework suggested in this thesis offers practical implications for identifying and addressing toxic speech to create a safer and more positive online environment. It does so by utilizing a speech-to-text algorithm in conjunction with a toxicity detection API. In order to improve the accuracy and generalizability of findings, future research should aim to address the limitations by incorporating more diverse datasets, investigating alternative techniques, and incorporating manual labeling processes.

Researchers and platform moderators can create efficient strategies and interventions to lessen the negative effects of toxic speech on online communities by understanding the characteristics of toxic speech, its distribution across topics, and its relationship with sentiment. Collaboration between researchers, platform creators, and users is necessary for ongoing research, policy development, and the promotion of positive and respectful interactions on TikTok and other similar platforms.

In conclusion, this study adds to the growing body of research on toxic speech in video-based platforms and serves as a foundation for future investigations and initiatives aimed at making online spaces safer and more welcoming.

References

Chen, E., Deb, A., & Ferrara, E. (2022). #Election2020: the first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science*, *5*(1), 1–18. https://doi.org/10.1007/s42001021001179

D'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. *2020 International Multi-Conference On: "Organization of Knowledge and Advanced Technologies" (OCTA)*, 1–5. https://doi.org/10.1109/OCTA49274.2020.9151853

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., & Meignier, S. (2018). An OpenSource Speaker Gender Detection Framework for Monitoring Gender Equality. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5214–5218. https://doi.org/10.1109/ICASSP.2018.8461471

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday*, *25*(11). https://doi.org/10.5210/fm.v25i11.11431

Hernandez Urbano Jr., R., Uy Ajero, J., Legaspi Angeles, A., Hacar Quintos, M. N., Regalado Imperial, J. M., & Llabanes Rodriguez, R. (2021, August 21). A BERT-based Hate Speech Classifier from Transcribed Online Short-Form Videos. *2021 5th International Conference on E-Society, E-Education and E-Technology*. https://doi.org/10.1145/3485768.3485806

Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Jigsaw. (2023). *Perspective API - How it works*. Perspectiveapi.com. https://perspectiveapi.com/how-it-works/

Matamoros-Fernandez, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television and New Media*, *22*(2), 205–224. https://eprints.qut.edu.au/207580/

Medina Serrano, J. C., Papakyriakopoulos, O., & Hegelich, S. (2020). Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok. *12th ACM Conference on Web Science*, 257–266. https://doi.org/10.1145/3394231.3397916

OpenAI. (2022, October 9). *Whisper*. GitHub. https://github.com/openai/whisper

Peeters, S. (2023, April 3). *Zeeschuimer (v1.5.0)*. Zenodo. https://zenodo.org/record/7796186

Peeters, S., & Hagen, S. (2022, September 28). *The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3914892

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Ilya Sutskever. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv. https://doi.org/10.48550/arxiv.2212.04356

Singh, N. K., Singh, P., & Chand, S. (2022). Deep learning based methods for cyberbullying detection on social media. *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 521–525. https://doi.org/10.1109/ICCCIS56430.2022.10037729

TikTok. (2022, December 8). *Countering hate on TikTok*. TikTok. https://www.tiktok.com/safety/en/countering-hate/

Vasconcellos, P. H. S., Lara, P. D. A., & Marques-Neto, H. T. (2023). Analyzing polarization and toxicity on political debate in brazilian TikTok videos transcriptions. *15th ACM Web Science Conference 2023 (WebSci '23)*, 33–42. https://doi.org/10.1145/3578503.3583613

Wu, C. S., & Bhandary, U. (2020). Detection of Hate Speech in Videos Using Machine Learning. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 585–590. https://doi.org/10.1109/CSCI51800.2020.00104

Ye, Y., Le, T., & Lee, D. (2023). *NoisyHate: Benchmarking Content Moderation Machine Learning Models with Human-Written Perturbations Online*. Conference acronym 'XX, Woodstock, NY. https://doi.org/10.48550/arxiv.2303.10430

Appendix

Figures

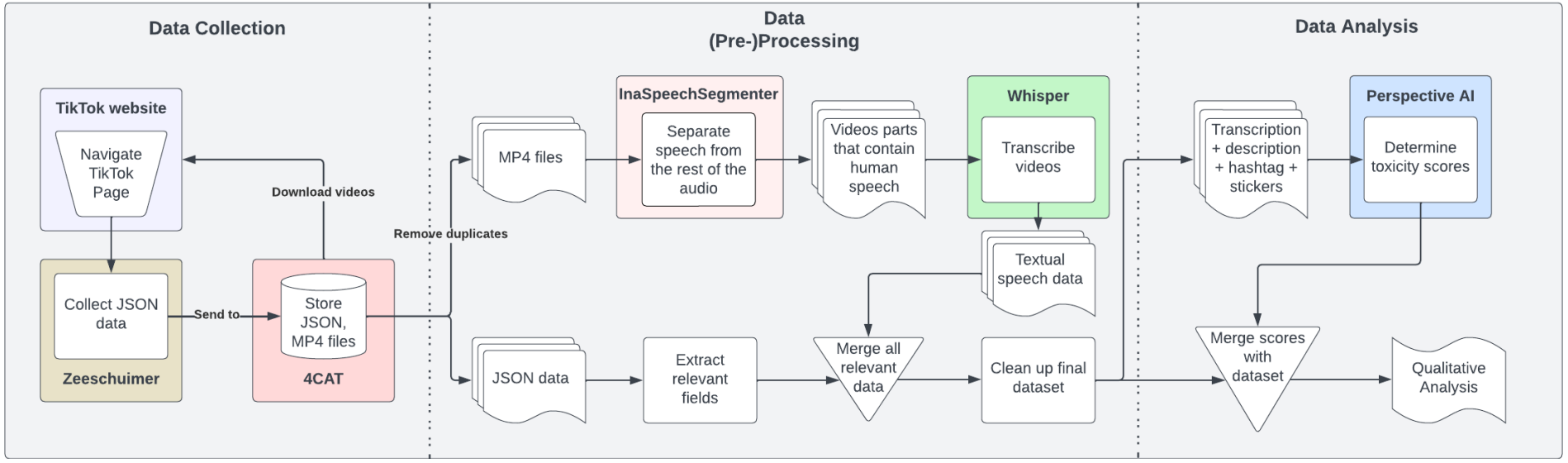

Figure 1 A thorough visual representation of the methodology used to produce this framework