Analyzing Polarization And Toxicity On Political Debate In Brazilian TikTok Videos Transcriptions

Paulo H. S. Vasconcellos
Pontifical Catholic University of Minas Gerais – PUC Minas
Belo Horizonte, Minas Gerais, Brazil
phsamus@gmail.com

Pedro Diógenes A. Lara
Pontifical Catholic University of Minas Gerais – PUC Minas
Belo Horizonte, Minas Gerais, Brazil
pedro.lara@sga.pucminas.br

Humberto T. Marques-Neto
Pontifical Catholic University of Minas Gerais – PUC Minas
Belo Horizonte, Minas Gerais, Brazil
humberto@pucminas.br

ABSTRACT

With the rise of TikTok's popularity, there is an opportunity to understand how political communication has been made on this platform based on short videos. In addition to understanding what topics and themes are discussed on TikTok, we also analyzed the polarization and toxic behavior of its users. However, this great opportunity brings a challenge as well. As TikTok is a video platform, the largest content information is in the video itself, so it is necessary to extract this information from video data and features.

In this paper, we propose a methodology to extract topics from TikTok video transcriptions in order to identify polarization and toxicity in their contents, by using techniques that range from web crawling to speech recognition algorithms. By providing a robust audio cleaning pipeline, it's possible to generate a less noisy dataset, by removing silence and music segments. We validate our methodology by practically applying it to create topics in order to identify signs of political polarization and toxicity in 8,329 Brazilian political TikTok videos, collected over the last two years. Our work shows that it is possible to extract coherent and meaningful topics from TikTok videos, even with the challenges spoken texts bring. We point out that topics related to religion and social classes contain a higher percentage of toxicity and polarization, as well as opposite hashtags, such as "direita" (Right-wing) and "esquerda" (Left-wing).

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; • **Networks** → *Online social networks.*

KEYWORDS

Political Polarization, Political Toxicity, Online Social Networks, Text Analysis, Topic Modeling, TikTok

ACM Reference Format:

Paulo H. S. Vasconcellos, Pedro Diógenes A. Lara, and Humberto T. Marques-Neto. 2023. Analyzing Polarization And Toxicity On Political Debate In Brazilian TikTok Videos Transcriptions. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3578503.3583613

WebSci '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583613

1 INTRODUCTION

For many years social media platforms have been used as political communication tools. Activists and political groups have seen these platforms as fruitful fields to communicate public policies [18] and social-economic events [40], but also, through some communities, to disseminate ideologies, polarization [17, 23], misinformation [16], and toxicity [6]. Although online social networks try to apply policies and security mechanisms to mitigate the risk of polarization and toxicity in the platform, literature has shown that this problem still persists. It's fair to expect this kind of social movement to repeat itself in newer social media platforms that quickly rise in popularity, such as TikTok.

In 2018, the Chinese company ByteDance released its social media platform named TikTok. In just a year, it became the world's most downloaded app [1] and, currently, contains over 1 billion users. We still know very little about political communication and activism on TikTok [28], especially about toxicity and polarization, and given its massive growth, this knowledge becomes necessary. One way to reach this goal is to model the content shared in the platform as topics and try to obtain meaning from them. Given that the nature of the platform's published content is video content instead of text like on Facebook and Twitter, it becomes necessary to adapt techniques of natural language processing and artificial intelligence for this distinct context.

In this paper, we propose a methodology to extract topics from TikTok video transcriptions in order to identify polarization and toxicity in their contents, by using techniques that range from web crawling to state-of-the-art speech recognition algorithms to extract textual information from TikTok videos. We apply our methodology to 8,329 Brazilian political TikTok videos created over the last two years to identify toxic content and polarization in the political debate. We show that, by using our methodology, it's possible to extract meaningful and coherent topics from TikTok videos. Our results indicate signs of polarization and toxicity in Brazilian political TikTok videos.

This work is organized in the following manner: In the next section, we show related work, and after that, we discuss the methodology adopted for this work, explaining each step shown in Figure 1. In the next sections, we apply our methodology, modeling the videos with BTM, and we present the results in Section 5, including an analysis of toxicity and polarization in Brazilian political video. Finally, we point out some ethical concerns around this contribution in Section 6 and discuss future work in Section 7.

[1] https://sensortower.com/blog/tiktok-revenue-downloads-2019

2 RELATED WORK

In Sections 2.1, 2.2 and 2.3 we describe some of the contributions around (i) polarization and toxicity in political debate, and the usage of (ii) topic modeling and (iii) NLP for modeling and analysis of the political debate on social media.

2.1 Polarization and Toxicity In Political Debate

Polarization and toxicity in online social networks for political debate is a widely researched field, especially in social networks such as Twitter [10, 17, 26, 38, 40, 44]. Besides that, it is important to understand that polarization is found in different social and political spheres, where its presence is found in contexts ranging from public policies (e.g. vaccination [23]) to ideology conflicts (e.g. abortion). One of the main concerns related to social networks is the fact that these platforms might potentialize the creation of "echo chambers" [20]: an effect that happens when a person experiences a biased, tailored media content that eliminates opposing viewpoints and differing voices. Cinelli et at. [9] observed this effect on many different online social networks, such as Twitter, Facebook, and Reddit.

2.2 Natural Language Processing Techniques For Online Social Media

Natural Language Processing is a sub-discipline of Computer Science, that involves Artificial Intelligence and Linguistics. It focuses on the generation and comprehension of human language. One of its greatest challenges is making so the way humans communicate with each other intelligible to a machine, with the goal of extracting meaning and language structure. Previous research applied one or more Natural Language Processing techniques, such as sentiment analysis [11], text mining [24], document similarity, semantic analysis [26], and word embeddings.

To generate descriptive statistics and analysis, authors have used text mining techniques that allow them to create visualizations such as word clouds and term frequency numbers. Other tasks involved the identification of similar documents, such as the work of Pita et al. [33], who applied document similarity techniques to identify politicians from Ecuador that had similar connections on Twitter.

Also, Automatic Speech Recognition (ASR) has been used along with NLP to create clusters and topics in transcripted videos, not only in political debate videos [28] but also in other contexts [2, 7]. However, most of the previous contributions didn't apply a well-defined pipeline structure to clean and manipulate the audio before submitting it to transcription algorithms, in order to remove noisy data. In this work, we use natural language processing techniques not only for pre-processing and visualization but also for topic modeling itself.

2.3 Topic Modeling of Political Debate

Topic modeling in social media platforms isn't a new subject, with studies already conducted in supervised and unsupervised learning techniques [15]. Platforms like Facebook and Twitter have seen thorough exploration in this field for different subjects, with politics not being excluded from them [8, 11, 37]. However, few contributions have been made to TikTok regarding the political sphere. Medina et al. [28] had one of the most outstanding contributions, analyzing duet videos (when users create response-type videos in split screen mode, displaying both videos simultaneously). They analyzed videos from two political hashtags, *#democrats* e *#republicans*, representing the two most popular American political parties. Their work focused exclusively on users from the USA and in two hashtags. As far as we know, no other contributions were made applying a methodology to explore political communication, polarization, and toxicity in Brazilian TikTok users.

We emphasize that topic modeling was used to identify many themes, from political topics [22] to agendas discussed and defended by politicians [21]. One of the most common techniques for topic modeling is Latent Dirichlet Allocation (LDA), with many papers adopting it as their main choice for topic generation [5, 15, 29, 34, 38]. LDA assumes that documents are composed of one of many topics and that each topic has many words attributed to it. Based on the words that form a text document, LDA calculates which topics may have taken part in its generation. If a document uses an abundance of words relevant to a determined topic, we may assume that this document belongs to that topic. It is noteworthy that the words composing each topic don't have the same probabilistic distribution. For instance, the word "COVID" has a higher chance of belonging to the topic "Pandemic," while the word "medicine" may have a lower chance since it may also belong to other topics.

However, there are some contributions that use alternatives to LDA, such as Jackson et al. [22], who used an unsupervised approach and still managed to efficiently extract topics by using a lexical approach, while Iyer et al. [21], on the other hand, used community detection techniques to group similar terms. Also, Xiaohui et al. proposed a novel algorithm focused on short documents and corpus called Biterms (BTM) [43]. Unlike classical approaches like LDA, which evaluates each document as a group of topics, Biterms models the entire corpus as a group of topics, making the inference process easier. After that, it assumes each Biterms (co-occurrence of a pair of words in a text interval) belongs to a topic, further reducing the complexity of the topic inference process by not inferring from each word. Finally, Grootendorst [14] created BERTopic: a topic modeling technique built upon Transformers architecture proposed by Vaswani et al. [41].

3 METHODOLOGY

TikTok's notoriety attracts different kinds of people, such as activists, organizations, and politically engaged people. Political contexts are excellent study subjects for topic modeling since the discussed topics may vary with time, showing even seasonal effects, like during elections. Furthermore, TikTok's videos are generally short-duration videos that may contain spoken language, differentiating it from other social media platforms where most content is created through written text.

Figure 1 shows the methodology we propose for this work. In the next sections, we describe each step of the methodology, providing more details on how they are executed.

3.1 Data Collection

Unlike other social media platforms, like Facebook or Twitter, TikTok does not yet offer an official API or third-party integration tools for data extraction. Also, unlike platforms like YouTube, TikTok's

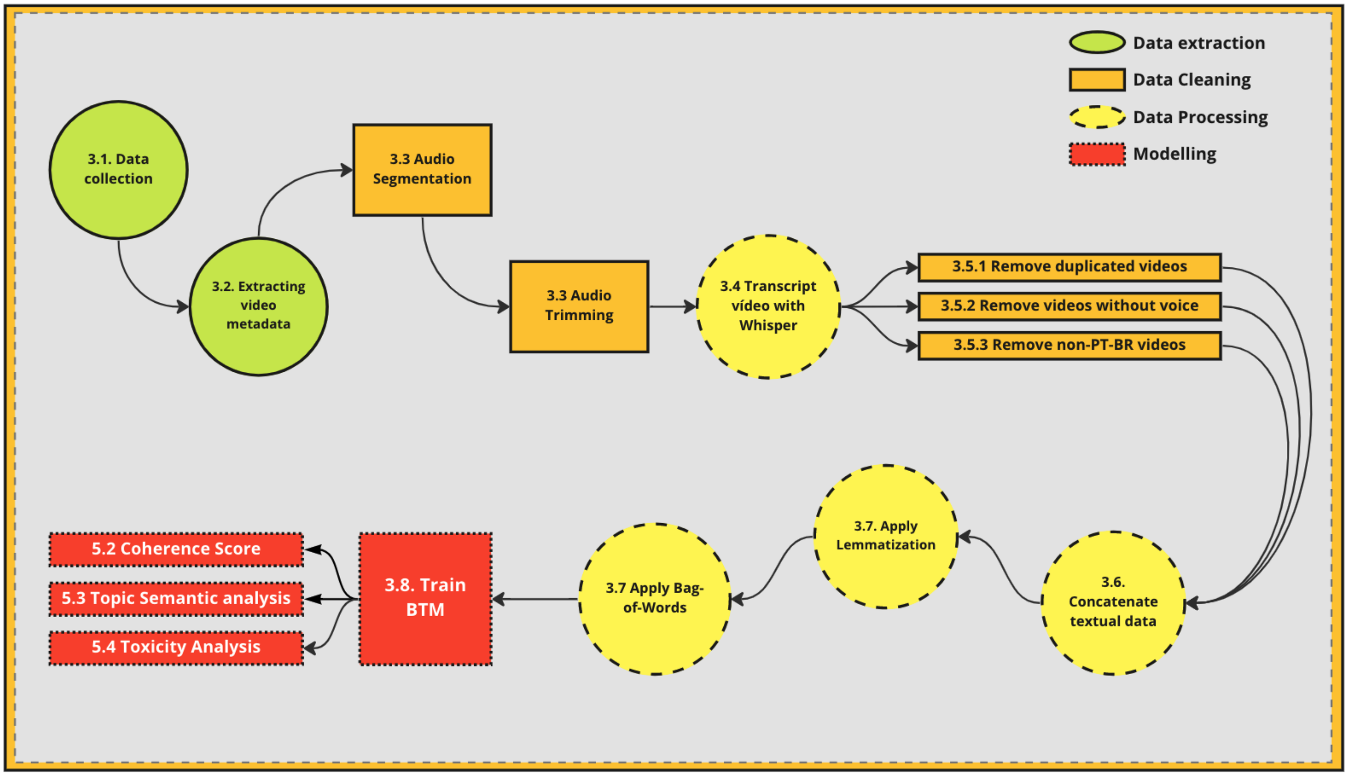

Figure 1: Proposed methodology, numbered as section numbers

own search features are limited since it doesn't allow its users to filter video characteristics or metadata, like dates, locale, and duration. TikTok does, however, allow users to search for hashtags. To fulfill this work's objectives, we used a crawler that searches videos based on hashtags and returns data and metadata for all these videos, such as description, all hashtags, likes and sharing statistics, and even a downloadable video link. We show the hashtags used by our crawler to collect the videos for our practical application in Table 1.

As our crawling solution, we used the TikTokAPI library [39], created by David Teather. It is a non-official API for TikTok which works by crawling the platform through its search page. All data and metadata returned by TikTokAPI is documented[2]. The crawler works through the results of TikTok's search engine, which doesn't offer filters or ordering features. As such, there are no criteria to determine the relevance, engagement, or recency of the search results returned by TikTok platform.

3.2 Extracting Video Metadata

One of the metadata returned by the crawler includes texts inserted through stickers. Stickers are elements that can be overlaid on the video during its uploads, such as GIFs, Emojis, and text. A lot of these sticker texts add some information to the video itself, that can be useful for topic modeling. We extract the texts from these stickers and concatenate them to the speech-to-text transcription results.

3.3 Audio Segmentation and Trimming

After we collected the videos, we needed to inspect their audio to understand which ones were relevant to this paper. A video is considered relevant when they have a person speaking in the video since it will help us to identify and model her speech. TikTok is well known for its dancing, viral, and satirical videos. Some of these videos might add little to none to discovering the political topic present in them, as many of them are only music and/or background sound with static images. This challenge becomes harder when the video itself is a mixture of the human voice, background music, and silence. An approach is necessary to not only identify if there is a person speaking but also remove useless audio segments, such as ones containing noise, silence, or only-music segments.

To tackle this problem, we choose to use the approach proposed by Doukhan et al. [12], who created a CNN-based audio segmentation toolkit that allows researchers to detect speech, music, noise, and even the speaker's gender from audio. We used an implementation of this solution, called the inaSpeechSegmenter[3] to separate the audio into several different segments. Figure 2 shows an illustration of the output of this step for a given audio, where inaSpeechSegments segments the audio into categories, such as general speech (when there is a person speaking, regardless of her gender), male speech, female speech, music, and noise (used when there is no speech or music), returning the audio file and the segmentation

[2]https://gist.github.com/davidteather/7c30780bbc30772ba11ec9e0b909e99d

[3]https://github.com/ina-foss/inaSpeechSegmenter

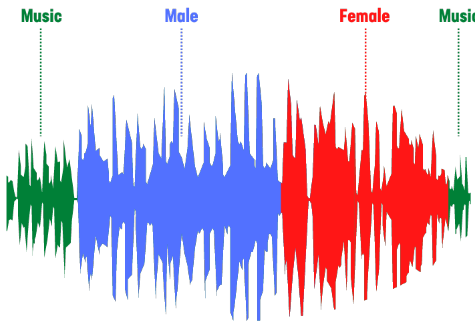

Figure 2: Illustration of the output provided by inaSpeechSegmenter. The framework provides the segments' time span and their label

divided by time spans. All videos collected in the data collection step by the crawler are passed in this framework.

After we segmented all audio files, we needed to extract the segments that contained any speech and remove music and noise segments. That's because we are only interested in human speech, regardless of gender, since this will be the input passed to the audio transcription step explained in Section 3.4. To manipulate the audio, we used the PyDub[4] library, which contains features like audio trimming. Using the segmentation time spans and labels, we removed every segment that didn't contain speech, as the only thing needed for the further steps of the methodology was the spoken part of the audio. Since TikTok is known for its dancing videos, many of the collected videos had only music in their content: a strategy adopted by users to manipulate TikTok's recommendation system, an effect observed by Klug et al. [25]. Music and song lyrics may reduce the performance of the topic modeling algorithm, so despite some songs having a political theme (e.g. politician jingles), they were still removed. The output of this step is a new file audio containing only speech segments.

3.4 Audio Transcription Using Whisper

TikTok, like other social media platforms, has textual elements in its content, such as the video's description. However, most part of the information is in the video itself through audio, so it is necessary to find a way to extract this data as textual information. We used speech-to-text algorithms to transcribe the video's audio content into text.

To obtain high-quality transcriptions, we used OpenAI's Whisper[5], a recent launched algorithm that has human-level performance in audio transcription. Different from other approaches for speech recognition, Whisper has a high level of quality in non-English audio, since one-third of its training dataset is composed of non-English data points, being robust enough to transcribe audio independently of the speakers' accent, regionalism, and voice speed. Also, Whisper is able to identify the language spoken in audio that it transcribes. This output will be important for our methodology since in Section 3.5 we filter non-Portuguese videos. The audio transcription will be the data used for the topic modeling task, along with other textual information described in Section 3.2. We used the official Python implementation of Whisper, open-sourced by OpenAI[6].

[4]https://github.com/jiaaro/pydub
[5]https://cdn.openai.com/papers/whisper.pdf

3.5 Video Filtering

To remove useless videos we used in our final dataset, we executed some filtering steps, such as:

3.5.1 Remove duplicated videos. Some videos might return twice if they are using two or more hashtags chosen for the crawler. For instance, a video that contains both hashtags "leftiktok" and "rightiktok" might be returned twice if the researcher used both tags in the web crawling step.

3.5.2 Remove videos without voice. In this step, we removed videos that do not contain a spoken voice. We know that we might lose a bit of information by removing such videos containing only background music, such as when a user declares support for one candidate by writing the candidate's name in the video. However, videos containing only music with static images are video types that may pollute the analysis of the generated topics and even make the results biased since it is difficult to identify their underlying themes. Through the usage of the audio segmentation library previously mentioned in Section 3.3, we were able to remove videos that contained only "music" or "noise" labels.

3.5.3 Remove non-PT-BR videos. Many videos returned by the crawler contained foreign languages. For instance, the hashtag "pt" is frequently present in Spanish language videos since it means "Para Ti" (or "For You," a hashtag used by TikTok users as they believe it may boost the odds of their video being recommended to other users); the hashtag "Lula" seems to be used in Thai videos about a local influencer; and, finally, hashtags about political movements that started in USA, such as "lefttiktok" or "righttiktok" also return videos in English and Italian languages. To deal with this, we used the language detected by the Whisper AI to remove videos made in a non-Portuguese language.

3.6 Data pre-processing

The transcribed texts needed cleanups and processing to enable the extraction of the highest amount of information from the political context. Given that the task's base data is text transcribed from videos, one of the biggest challenges we found was the need to identify and remove fillers[7]: words and sounds that people use during conversation to indicate that they need to think but haven't finished speaking. In American English, the most common filler sounds are *ah* or *uh* and *um* [4]. Examples of filler words in Brazilian Portuguese include *é* ("is"), *hum* ("um"), *então* ("so"), *tipo* ("like"), *bem* ("well"), *né*, and many others. Written text tasks don't have such issues since we don't use fillers in written text. These are speech elements that end up transferring to text after transcription. The full list of filler words removed is available on our code repository.

Finally, for each video, we concatenate all the textual information we extracted into one string that will be used in all further steps.

[6]https://github.com/openai/whisper
[7]https://en.wikipedia.org/wiki/Filler_(linguistics)

The information in each string is the video's description returned by the crawler, the audio content transcripted by Whisper, the video's hashtags, and the "sticker" texts embedded into the video.

3.7 Lemmatization and Word Embedding

The next step is to prepare the strings for the topic modeling task. We use lemmatization and word embedding techniques for this goal. Lemmatization is one of the most common text pre-processing techniques for Natural Language Processing. It is the process of removing variations and inflected forms of a word, reducing it to its dictionary form, called "lemma." It differs from the Stemming technique in that the word's context is kept during lemmatization. By applying this technique, we can group several words with similar meanings into one word, reducing variation during topic modeling. The usage of lemmatization over stemming provides better morphological analysis and can be beneficial for languages with gender variations of words, such as "rei" and "rainha" (king and queen in Portuguese, respectively). We used the library Spacy, by Explosion AI, which provides an accurate lemmatizer for English and other languages, such as Portuguese[8]. As its training dataset, Spacy uses UD Portuguese Bosque [35] and WikiNER [32] for the Portuguese lemmatizer.

3.8 Topic Modeling With BTM

Our methodology proposes the usage of Biterms [43] as the algorithm to model TikTok video data as topics. This is a topic modeling problem and currently, there are good options for algorithms, such as the already mentioned in Section 2 LDA [3] and BERTopic [14], as well as other lexical approaches to the problem [22]. However, we want our methodology to be scalable and flexible enough to model topics from different contexts and situations, such as small data sets. Some hashtags are underused, niched, or less popular, which might return just a few and/or short videos in the search results. So, a less complex and robust algorithm must be selected to model to include short documents and small datasets among the possibilities. As such, our technique of choice for this methodology is BTM, due to its flexibility, simplicity, and robustness.

4 EXPERIMENTAL SETUP

As one of the objectives of our work, we applied the proposed methodology to a practical problem. In this Section, we present more details about the experiment that collected Brazilian political TikTok videos and modeled topics from its textual content. We consider a Brazilian political TikTok video the one that contains one or more hashtags presented in Table 1. In Section 5 we show the results of this experiment that aims to identify signs of toxicity and polarization on TikTok.

4.1 Data Overview

This section shows some descriptive statistics of the dataset used to apply the proposed methodology. Table 1 show the hashtags used to collect our dataset using the TikTokApi. The hashtags were chosen based on (i) the three most well-positioned presidential candidates in the vote surveys at the time (Lula, Bolsonaro, and Ciro Gomes), (ii) hashtags that reflect the political spectrum [42] such as "direita" (Right-wing), "esquerda" (Left-wing), "politica" (politics), and "politicabrasileira" (Brazilian politics); and (iii) hashtags about political movements created on TikTok, such as #leftiktok and #rightiktok. We avoided using hashtags such as #bolsonaro22 and #lula13 in our analysis since it might lead to a more biased topic modeling. Generally, hashtags like #bolsonaro22 and #lula13 return pro-candidate videos, once those numbers (22 and 13) are the candidate's numbers that voters must inform on election day. In total, we collected 13,696 raw videos. After we pass them through our methodology, our final dataset contains 8,329 videos that will be used to model topics using BTM. There's a median of 103 tokens per video, and a median of 71 unique tokens per video. After applying the bag-of-words embedding, our vocabulary contains 1,000 terms. The videos were created between 2019-10-13 and 2022-09-25.

Table 1: Hashtags searched and amount of videos returned

Hashtag	# Videos
direita	1,505
esquerda	1,514
leftiktok	1,250
lefttiktok	881
rightiktok	346
righttiktok	346
lula	1,961
bolsonaro	2,799
ciro	262
cirogomes	482
politica	1,462
politicabrasileira	888

5 RESULTS

In this Section, we show some results of the application of the proposed methodology. The code used for all steps is available in our GitHub repository[9].

5.1 Co-occurence of Hashtags In Videos

Figure 3 shows the co-occurrence of the most used hashtags in TikTok videos in our dataset. In short, the heatmap shows the top 25 most used hashtags and the number of videos that both appears together. To make it easier to visualize, we adjusted to 600 the max value (darker red) in the color scale; which means that cells higher than 600 videos will keep the same red tone. In this figure, we can see that 1066 videos that used the hashtag "bolsonaro" also used a "lula" hashtag; and 642 videos used both the hashtags "direita" (Right-wing) and "esquerda" (Left-wing). That suggests that users create political videos and want them to be found by people with different opinions from their own, however, a second-order effect that might be generated is the increase of polarization in the platform. Although using both "direita" and "esquerda" hashtags might be inoffensive, using extremely opposite hashtags might generate a polarized debate. We can see some examples in Figure 3, such as

[8] https://spacy.io/models/pt#pt_core_news_md

[9] https://github.com/paulozip/tiktok_polarization_toxicity_analysis

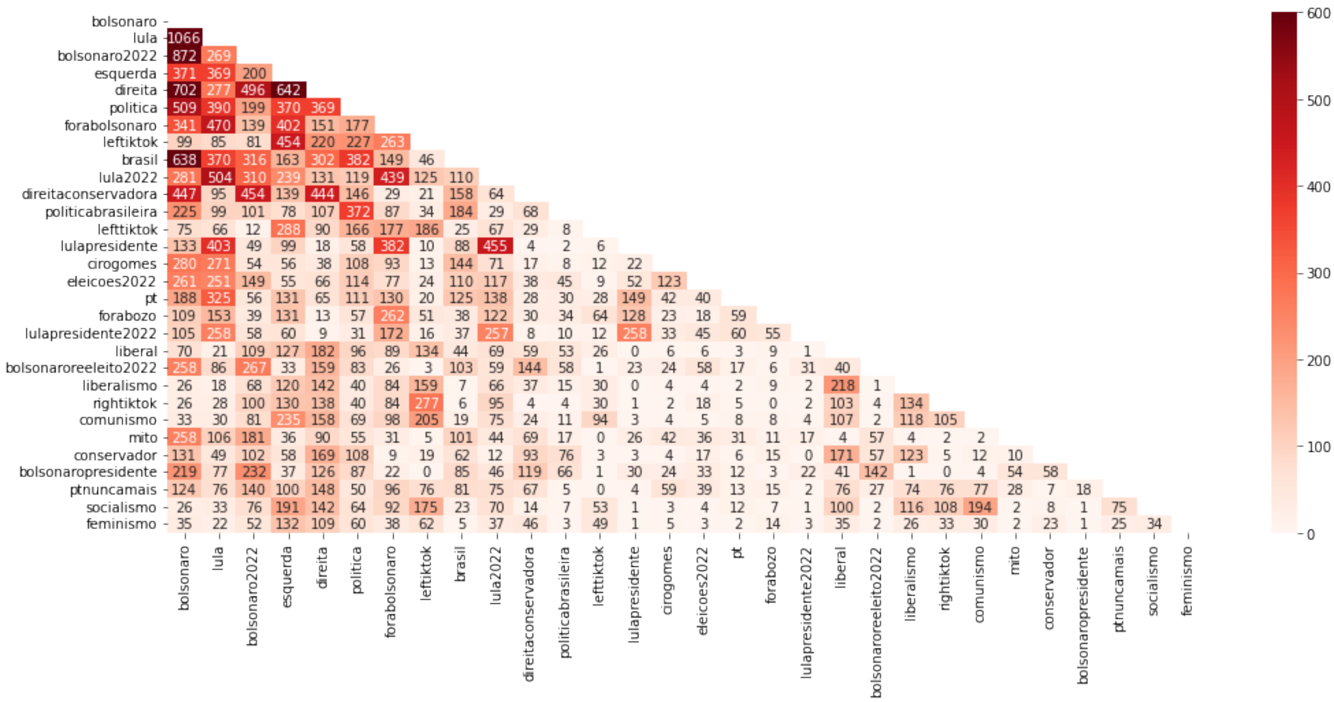

Figure 3: Co-occurrence of top 25 most hashtags used in TikTok videos

the hashtags "lulapresidente" (meaning "Lula for president") and "forabolsonaro" (a protest against Bolsonaro), or "ptnuncamais" (PT never again, where PT is Lula's party name) and "bolsonaro2022" (suggesting Bolsonaro's re-election). In a recent contribution, Klug et al. [25] show that users tend to believe that pilling up hashtags influences how the TikTok algorithm works, so using harmful hashtags might propagate the polarization in the platform. This thought is reinforced in Section 5.4, where we show that videos with such extreme hashtags contain, frequently, toxic behavior.

5.2 Coherence Score

There's a challenge in evaluating the quality of topic modeling algorithms, especially due to their unsupervised origins, where the lack of ground-truth data to estimate their real performance. However, there are good options to tackle this problem. To estimate the best number of topics, we used the coherence score as our qualitative metric.

Coherence measures the semantic similarity in each topic. A higher coherence score indicates a topic with documents semantically similar, while a lower score indicates higher chances of a topic containing random documents and not correlated. Once there is no standard definition of what is good or bad coherence score (e.g. for a given task, a coherence score of -3.00 might be good enough, while for other contexts it might be bad), it's well known that the main goal is to maximize such a score, which suggest a higher semantic value in the generated topics. The coherence score has been used as one of the main quantitative metrics to assess topic modeling performance for algorithms like BTM [30, 43]. We used UMass Coherence Score to identify the best number of topics to be

Table 2: Average, median, and standard deviation for *UMass Coherence score* for BTM, from 10 to 30 topics. Best results are highlighted in bold.

# Topics	Median	Mean	SD
10	-1.51	-1.54	0.26
11	-1.55	-1.52	0.18
12	**-1.42**	**-1.47**	0.24
13	-1.53	-1.58	0.33
14	-1.59	-1.57	0.13
15	-1.56	-1.55	0.23
16	-1.54	-1.49	0.26
17	-1.63	-1.58	0.25
18	-1.67	-1.65	0.24
19	-1.53	-1.56	0.31
20	-1.55	-1.56	0.25
21	-1.55	-1.53	0.33
22	-1.55	-1.59	0.32
23	-1.55	-1.61	0.38
24	-1.66	-1.64	0.39
25	-1.63	-1.67	0.35
26	-1.56	-1.60	0.30
27	-1.58	-1.64	0.32
28	-1.64	-1.66	0.37
29	-1.64	-1.69	0.41
30	-1.71	-1.71	0.36

used in Biterms. UMass Coherence Score calculates the frequency that two terms appear together in the corpus, where a value closer to 0 indicates a more coherent topic. Table 2 shows the coherence score for a given number of topics in BTM. For each number of topics, we assess the overall coherence using its mean, median, and standard deviation. Based on this table, the optimal number of topics for our dataset is 12, which was selected to model our dataset as topics.

5.3 Topic Semantic Analysis

After generating the 12 topics, we started analyzing them in a qualitative manner, in order to find semantic similarities between the top terms in each topic. For that step, we performed a semantic analysis in those topics using *relevance*, proposed by Sievert e Shirley [36], where the relevance of a term in a given topic is given by the probability log of such term in the topic, summed to the probability log of such term in the corpus, both of them weighted by a λ parameter. The output of this step can be seen in Table 3, where for each topic we show the top 10 most relevant terms, the number of documents (videos) assigned to such topic, and the percentage of the corpus that this number of documents represents. We also set a label for each topic, according to the top terms.

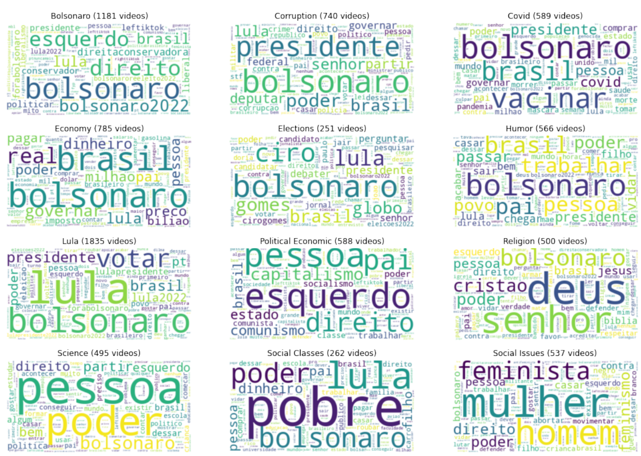

Figure 4: Word cloud of most frequent terms in each topic

We can see that the top terms in each topic are semantically similar. For example, we can see terms like "vacinar" (to vaccinate), "covid", "mascara" (mask), and "virus", which are correlated in videos about the pandemic and Covid-19. The same can be observed in topics such as "Political Economic", which contains videos related to political-economic systems, such as "capitalismo" (capitalism), "marx", "socialismo" (socialism), and so on. Also, BTM was able to group different types of social issues videos on the same topic, "Social Issues". This topic contains videos about feminism and LGBTQIA+, but also about sexism and racism. Finally, our model was able to generate topics related to the two most popular presidential candidates in the Brazilian elections: Lula and Bolsonaro. While the top terms in Lula topic are more pro-Lula, such as "lulalivre" (Free Lula movement) and "lulapresidente", the Bolsonaro topic contains pro-Bolsonaro terms and anti-Lula terms. This analysis might suggest that we have signs of polarization on TikTok. We explore more of this aspect in Section 5.4, where we conducted a toxicity analysis on each topic.

Along with an analysis of the most important terms, we can see in Figure 4 a word cloud of the most frequent terms present on each topic. We can see that the term "Bolsonaro" appears frequently in all topics, but besides that, many of the frequent terms that appear in the topics are coherent with the theme of the topic itself. For instance, in the Religion topic, terms like "Deus" (God), "cristão" (Christian), and Jesus are expected to appear together. The same can be seen in other topics, like "Social issues", with "Feminista" (Feminist) being one of the most frequent terms. By analyzing the results of both the word cloud in Figure 4 and the most important terms in Table 3 we can see that our methodology allowed us to create meaningful and coherent topics.

5.4 Polarization and Toxicity in TikTok Videos

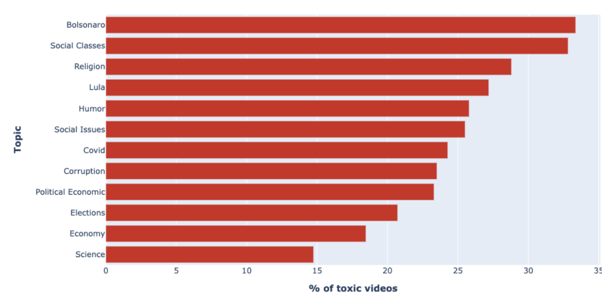

Figure 5: Percentage of toxic videos per topic

Political polarization might be defined as the divergence of political attitudes away from the center, which leads to ideological extremes [1]. Iyengar et al. [20], on other hand, define that an important aspect of polarization relies on the fact that one group dislikes the oppositive group or party. As online social networks have been widely researched due to their potential to create echo chambers, this polarized environment might lead to a more toxic platform [13, 31].

As aforementioned, videos that share opposite hashtags might indicate a polarized platform, and when combined with a toxicity analysis, might reinforce the definition proposed by Iyengar et al. Table 4 shows an analysis of the most frequent pair of hashtags used in each topic. We can see the presence of opposites hashtags in most topics, such as "leftiktok/rightiktok" and "bolsonaro/lula". When combined with the toxicity analysis in Figure 6, we can note some of these hashtags are one of the most toxic hashtags pair in our dataset. Also, we can observe that the topics "Science" and "Bolsonaro" does not contain opposite hashtags in Table 4, however, if we analyze the latter in Table 3, we can see that many of the most important terms identified by BTM can be understood as toxic and polarized terms, such as "foralula" (Out Lula), "forapt" (Out PT – Worker's Party), and "ptnuncamais" (PT Never Again).

Table 3: Overview topics generated by BTM

Topic	Top 10 terms	#docs	%corpus
Lula	lula, lulapresidente, turno, presidente2022, votar, 2022, lula13, lulalivre, impresso, 13	1835	22.03
Economy	petroleo, preco, dolar, gasolina, petrobras, combustivel, custar, gas, biliao, imposto	785	9.42
Social Issues	mulher, homem, feminista, feminismo, feminino, machista, movimentar, preto, genero, estuprar	537	6.44
Religion	biblia, jesus, cristao, deus, igreja, religiao, calmo, amor, cancelar, senhor	500	6.00
Political Economic	classe, capitalismo, capitalista, marx, producao, socialismo, propriedade, comunismo, socialista, coreia	588	7.05
Social Classes	pobre, faculdade, carro, universidade, escola, dinheiro, aviao, celular, roubar, bolsar	262	3.14
Humor	cantar, cafe, manha, chorar, sonhar, cheio, dormir, gritar, povo, horar	566	6.79
Corruption	tribunal, supremo, federal, investigar, senado, ministro, constituicao, stf, juiz, senador	740	8.88
Bolsonaro	foralula, forapt, bolsonarotemrazao, bolsonaropresidente, ptnuncamais, direita, bolsonaroreeleito, conservador, bolsonaro, conservadorismo	1181	14.18
Elections	gomes, ciro, doria, folha, globo, joao, candidato, jornal, jair, cirogomes	251	3.01
Covid	vacinar, covid, mascara, virus, cloroquina, morte, pandemia, russia, culpar, mundial	589	7.07
Science	ciencia, conteudo, artigo, conhecimento, utilizar, comum, assunto, formar, bastante, visto	495	5.94

Table 4: Top 5 most used hashtag pairs per topic

Topic Name	Top 1	Top 2	Top 3	Top 4	Top 5
Science	esquerda/leftiktok	liberal/rightiktok	liberalismo/rightiktok	liberal/liberalismo	conservadorismo/rightiktok
Bolsonaro	liberal/rightiktok	bolsonaro/bolsonaro2022	liberalismo/rightiktok	bolsonaro2022/direita	liberal/liberalismo
Lula	bolsonaro/lula	lula/lula2022	lula/lulapresidente	lula2022/lulapresidente	forabolsonaro/lula
Covid	forabolsonaro/leftiktok	esquerda/forabolsonaro	bolsonaro/brasil	esquerda/leftiktok	bolsonaro/direita
Social Issues	esquerda/feminismo	feminismo/leftiktok	direita/feminismo	feminismo/feminista	direita/esquerda
Social Classes	bolsonaro/lula	esquerda/leftiktok	direita/esquerda	lula/lulapresidente	bolsonaro/direita
Humor	bolsonaro/bolsonaro2022	bolsonaro/lula	direita/direitaconservadora	bolsonaro/direita	bolsonaro/brasil
Corruption	bolsonaro/bolsonaro2022	bolsonaro/lula	bolsonaro/direita	bolsonaro/brasil	politica/politicabrasileira
Political Economic	esquerda/leftiktok	comunismo/leftiktok	direita/esquerda	leftiktok/socialismo	leftiktok/rightiktok
Economy	bolsonaro/lula	leftiktok/rightiktok	esquerda/leftiktok	forabolsonaro/leftiktok	direita/esquerda
Religion	bolsonaro/bolsonaro2022	bolsonaro/brasil	bolsonaro/direita	bolsonaro/lula	bolsonaro2022/brasil
Elections	bolsonaro/lula	bolsonaro/cirogomes	cirogomes/lula	bolsonaro/ciro	bolsonaro/eleicoes2022

In order to analyze toxicity in the videos, a model that is capable to identify toxic content through its text is necessary. Despite having some good options to apply toxicity detection in English text, such as VADER, proposed by Hutto and Gilbert [19], there are few options to apply toxicity detection in Brazilian Portuguese. Fortunately, Leite et al. [27] created ToLD-Br: a Toxic Language Dataset for Brazilian Portuguese. ToLD-Br is a dataset used to fine-tune BERT-based models on Brazilian Portuguese data associated with toxic words related to racism, xenophobia, misogyny, and other contexts. A BERT-based model was fine-tuned for binary and multi-label classifications, allowing users to identify toxic texts. A video is considered toxic if a classification score is higher than the threshold of 0.5 on any label (e.g. racism). This pre-trained model was made available by the authors[10], and we used it to classify a video text into toxic or non-toxic.

Figure 5 shows a combined analysis by applying the toxicity detection model with the topic generated by Biterms. Our analysis shows that videos related to social classes and religion contain a higher level of toxicity. For social classes, there are several videos that toxically criticize social inequality (e.g. poverty), access to education (e.g. racial quotas), and employment. For religion, we have toxic debates about different aspects of religion in our society (e.g. feminism and Christianity), most of them criticizing religious intolerance and traditional values. Also, videos related to presidential candidates like Bolsonaro and Lula contain a higher level of toxicity, with almost 35% and 27% of the videos being toxic, respectively. This shows signs of polarization in the platform, where divergent points of view are aggressively attacked by their discordant [20]. And, different from other social networks, TikTok's features might increase polarization in the platform, such as Duet and Stitch. Duet allows users to have their videos playing split-screen or green-screened next to another user's video, while TikTok Stitch lets you incorporate up to 5 seconds of someone else's content into your own TikTok video.

[10]https://github.com/JAugusto97/ToLD-Br

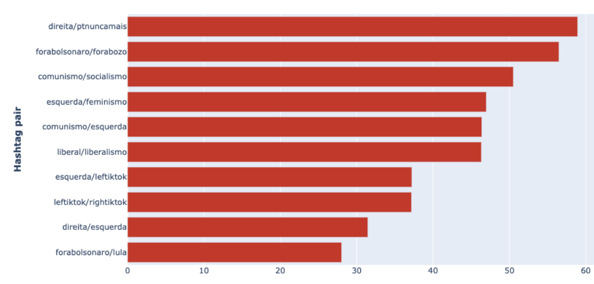

Figure 6: Percentage of toxic videos in the top 10 most used hashtag pairs

A similar effect can be seen when we apply the same analysis in the co-occurrence of hashtags shown in Figure 3. Figure 6 shows the top 10 most toxic pairs of hashtags used in videos. This analysis can help us to understand toxicity levels between different political views, which might lead to a higher polarization on TikTok. We can see in Figure 6 that almost 60% of videos containing both hashtags "direita" (Right-wing), and "ptnuncamais" (PT never again, where PT is a Left-wing party) are considered toxic. The opposite is also true, where around 27% percent of videos containing the hashtags "forabolsonaro" (as Bolsonaro is a Right-wing candidate) and "lula" (a Left-wing candidate) are toxic. These analyses reinforce the polarization definition made by Iyengar et al. [20], where aggressive and toxic videos are posted by both sides of the political spectrum.

6 LIMITATIONS, REPRODUCIBILITY AND ETHICAL CONCERNS

Some ethical questions must be taken into consideration when we make public a study like this one. One of them is related to the usage of web crawlers to collect data from TikTok. Although there's no formal definition of the legal aspect of a web crawler (with many cases ending in jurisprudence), it's important to take into consideration that public data collected from TikTok must be in conformity with the data protection laws and policies of each country. We want to make it easy for researchers to reproduce our work, but it's important to take data protection into account. To do so, we deleted the collected data after the analyses and application of this methodology. However, we preserved the video IDs to allow the reproducibility of this work. All IDs can be found in our repository[11].

Besides that, we identified two main threats to our work. The first is related to the methodology itself, in which the data collection process depends on an external tool responsible for crawling the platform. If TikTok applies a more strict policy, without providing an official resource to extract videos from the platform, it might difficult the data collection process, which is crucial for the methodology. The second one is related to hashtag selection, where a different set of hashtags might generate different topics and levels of toxicity and polarization. Despite that, we understand that the results of applying the proposed methodology contribute to the understanding of polarization and toxicity in the platform in a scalable and practical manner.

7 CONCLUSION AND FUTURE WORK

We presented a scalable and flexible methodology to identify toxicity and polarization in political debate on TikTok. We were able to model topics in transcripted TikTok videos to detect polarization and toxicity in Brazilian political videos, still, there are some opportunities to leverage this methodology, such as improvement in the toxicity method to detect sarcasm and pejorative words, since they may have a toxic discourse, but that is difficult to model. Also, this work can be expanded to understand gender equality in the political debate on TikTok, as our methodology includes the usage of audio segmentation that is able to differentiate male and female speech [12], as well as be used in contributions related to analyzing hate speech on TikTok.

ACKNOWLEDGMENTS

This work was supported by FAPEMIG (PPM- 00253-18), CNPq, and PUC Minas.

REFERENCES

[1] Delia Baldassarri and Andrew Gelman. 2008. Partisans without Constraint: Political Polarization and Trends in American Public Opinion. *Amer. J. Sociology* 114, 2 (2008), 408–446. https://doi.org/10.1086/590649 arXiv:https://doi.org/10.1086/590649

[2] Subhasree Basu, Yi Yu, and Roger Zimmermann. 2016. Fuzzy clustering of lecture videos based on topic modeling. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 1–6. https://doi.org/10.1109/CBMI.2016.7500264

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (mar 2003), 993–1022.

[4] Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44 (2001), 123 – 147.

[5] Rodrigo Castro and Carmen Vaca. 2017. National Leaders' Twitter Speech to Infer Political Leaning and Election Results in 2015 Venezuelan Parliamentary Elections. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 866–871. https://doi.org/10.1109/ICDMW.2017.118

[6] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (Troy, New York, USA) *(WebSci '17)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3091478.3091487

[7] Berlin Chen. 2009. Word Topic Models for Spoken Document Retrieval and Transcription. *ACM Transactions on Asian Language Information Processing* 8, 1, Article 2 (mar 2009), 27 pages. https://doi.org/10.1145/1482343.1482345

[8] Shu-I Chiu and Kuo-Wei Hsu. 2018. Predicting Political Tendency of Posts on Facebook. In *Proceedings of the 2018 7th International Conference on Software and Computer Applications* (Kuantan, Malaysia) *(ICSCA 2018)*. Association for Computing Machinery, New York, NY, USA, 110–114. https://doi.org/10.1145/3185089.3185094

[9] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118. https://doi.org/10.1073/pnas.2023301118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2023301118

[10] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. *Fifth International AAAI Conference on Weblogs and Social Media.*

[11] Saurabh Dorle and Nitin Pise. 2018. Political Sentiment Analysis through Social Media. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*. 869–873. https://doi.org/10.1109/ICCMC.2018.8487879

[12] David Doukhan, Jean Carrive, Felicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5214–5218. https://doi.org/10.1109/ICASSP.2018.8461471

[13] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 913–922. https://doi.org/10.1145/3178876.3186139

[14] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[15] Libby Hemphill and Angela M. Schöpke-Gonzalez. 2019. Two Computational Models for Analyzing Political Attention in Social Media. arXiv:1909.08189 [cs.SI]

[16] Brian Heredia, Joseph D. Prusa, and Taghi M. Khoshgoftaar. 2018. The Impact of Malicious Accounts on Political Tweet Sentiment. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. 197–202. https://doi.org/10.1109/CIC.2018.00035

[17] Sameera Horawalavithana, Kin Wai Ng, and Adriana Iamnitchi. 2021. Drivers of Polarized Discussions on Twitter during Venezuela Political Crisis. In *13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 205–214. https://doi.org/10.1145/3447535.3462496

[18] Rocío B. Hubert, Elsa Estevez, Ana Maguitman, and Tomasz Janowski. 2020. Analyzing and Visualizing Government-Citizen Interactions on Twitter to Support Public Policy-Making. *Digit. Gov.: Res. Pract.* 1, 2, Article 15 (apr 2020), 20 pages. https://doi.org/10.1145/3360001

[19] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[20] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. 2012. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly* 76, 3 (09 2012), 405–431. https://doi.org/10.1093/poq/nfs038 arXiv:https://academic.oup.com/poq/article-pdf/76/3/405/17120500/nfs038.pdf

[21] Rohit Iyer, Johnny Wong, Wallapak Tavanapong, and David A. M. Peterson. 2017. Identifying Policy Agenda Sub-Topics in Political Tweets Based on Community Detection *(ASONAM '17)*. Association for Computing Machinery, New York, NY, USA, 698–705. https://doi.org/10.1145/3110025.3116208

[11] https://github.com/paulozip/tiktok_polarization_toxicity_analysis

[22] Sam Jackson, Feifei Zhang, Olga Boichak, Lauren Bryant, Yingya Li, Jeff Hemsley, Jennifer Stromer-Galley, Bryan Semaan, and Nancy McCracken. 2017. Identifying Political Topics in Social Media Messages: A Lexicon-Based Approach. In *Proceedings of the 8th International Conference on Social Media; Society* (Toronto, ON, Canada) *(#SMSociety17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 10 pages. https://doi.org/10.1145/3097286.3097298

[23] Elise Jing and Yong-Yeol Ahn. 2021. Characterizing partisan political narrative frameworks about COVID-19 on Twitter. *EPJ Data Sci.* 10, 1 (2021), 53. https://doi.org/10.1140/epjds/s13688-021-00308-4

[24] Jufri, Aedah Binti Abd. Rahman, and Suarga. 2020. The Framework for Political Communication Text Mining Based on Twitter. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*. 1–6. https://doi.org/10.1109/ICORIS50180.2020.9320805

[25] Daniel Klug, Yiluo Qin, Morgan Evans, and Geoff Kaufman. 2021. Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. In *13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 84–92. https://doi.org/10.1145/3447535.3462512

[26] Mirko Lai, Cristina Bosco, Viviana Patti, and Daniela Virone. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 1–9. https://doi.org/10.1109/DSAA.2015.7344884

[27] João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 914–924. https://aclanthology.org/2020.aacl-main.91

[28] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok. *12th ACM Conference on Web Science* (Jun 2020). https://doi.org/10.1145/3394231.3397916

[29] Armin Mertens, Franziska Pradel, Ayjeren Rozyjumayeva, and Jens Wäckerle. 2019. *As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians*. Association for Computing Machinery, New York, NY, USA, 193–201. https://doi.org/10.1145/3292522.3326013

[30] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, United Kingdom) *(EMNLP '11)*. Association for Computational Linguistics, USA, 262–272.

[31] Gianmarco Morales, Corrado Monti, and Michele Starnini. 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports* 11 (02 2021). https://doi.org/10.1038/s41598-021-81531-x

[32] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2017. Learning multilingual named entity recognition from Wikipedia. (10 2017). https://doi.org/10.6084/m9.figshare.5462500.v1

[33] Oswaldo Pita, Gabriela Baquerizo, Carmen Vaca, Jonathan Mendieta, Mónica Villavicencio, and Jorge Rodríguez. 2016. Linguistic profiles on microblogging platforms to characterize political leaders: The Ecuadorian case on Twitter. In *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*. 1–6. https://doi.org/10.1109/ETCM.2016.7750870

[34] Adnan Qayyum, Zafar Gilani, Siddique Latif, and Junaid Qadir. 2018. Exploring Media Bias and Toxicity in South Asian Political Discourse. In *2018 12th International Conference on Open Source Systems and Technologies (ICOSST)*. 01–08. https://doi.org/10.1109/ICOSST.2018.8632183

[35] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*. Pisa, Italy, 197–206. http://aclweb.org/anthology/W17-6523

[36] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics, Baltimore, Maryland, USA, 63–70. https://doi.org/10.3115/v1/W14-3110

[37] Deepak Soni, Mayank Sharma, and Sunil Kumar Khatri. 2017. Political opinion mining using E-social network data. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*. 163–165. https://doi.org/10.1109/ICTUS.2017.8285997

[38] Hiroki Takikawa and Kikuko Nagayoshi. 2017. Political polarization in social media: Analysis of the "Twitter political field" in Japan. In *2017 IEEE International Conference on Big Data (Big Data)*. 3143–3150. https://doi.org/10.1109/BigData.2017.8258291

[39] David Teather. 2021. *TikTokAPI*. https://doi.org/10.5281/zenodo.5773718

[40] James Usher, Lucía Morales, and Pierpaolo Dondio. 2019. BREXIT: A Granger Causality of Twitter Political Polarisation on the FTSE 100 Index and the Pound. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 51–54. https://doi.org/10.1109/AIKE.2019.00017

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[42] Wikipedia. 2022. Political spectrum. https://en.wikipedia.org/wiki/Political_spectrum

[43] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A Biterm Topic Model for Short Texts. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1445–1456. https://doi.org/10.1145/2488388.2488514

[44] Sarita Yardi and danah boyd. 2010. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society* 30 (10 2010), 316–327. https://doi.org/10.1177/0270467610380011