Topic Shift como um Proxy para Avaliar Politização nas Eleições Brasileiras de 2022

Marcelo Sartori Locatelli[1], **Virgilio Almeida** [1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Caixa Postal 702 – 31.270-901 – Belo Horizonte – MG – Brasil

locatellimarcelo@dcc.ufmg.br, virgilio@dcc.ufmg.br

Abstract. *Politicization is a social phenomenon characterized by the extent to which non-political topics are given a political tone. Over the years, topics such as religion, vaccines, and climate change were subject to intense politicization, especially in social media platforms, which enable the study of this process as massive amounts of data are available. We leverage such data to assess politicization by using a method based on topic shifts to or from politics. For this, we train a classifier using PU-learning, as political labels may be easily obtained from keywords while non-political labels are scarce. Our findings suggest that the studied platforms show evidence of politicization, especially in the discussion of more controversial topics such as economy and education.*

Resumo. *A politização é um fenômeno social caracterizado pelo grau de que tópicos não políticos recebem um tom político. Ao longo dos anos, tópicos como como religião, vacinas e mudanças climáticas foram sujeitos a intensa politização, especialmente nas redes sociais, o que permite o estudo desse processo, pois há grandes quantidades de dados disponíveis. Utilizamos esses dados para avaliar a politização por meio de um método baseado topic shifts de ou para a política. Para isso, treinamos um classificador usando PU-learning, pois os rótulos políticos podem ser facilmente obtidos a partir de palavras-chave, enquanto os não políticos são escassos. Nossos resultados sugerem que as plataformas estudadas mostram evidências de politização, especialmente na discussão de tópicos controversos, como economia e educação.*

1. Introdução

Com a expansão das redes sociais, qualquer pessoa pode compartilhar suas opiniões sobre um dado assunto com um alcance muito superior ao que seria possível sem essas tecnologias. Isso, no contexto de eleições, por exemplo, poderá levar a comentários e discussões políticas e polarização, a medida que os usuários começam a interagir mais com os conteúdos mais alinhados aos seus valores [Spohr 2017]. Esse fenômeno já muito estudado na literatura refere-se ao processo pelo qual dois ou mais grupos políticos, ao escolherem seletivamente consumir opiniões com as quais já concordam, adotam pontos de vista cada vez mais extremos e antagonistas, criando as chamadas bolhas [Layton et al. 2021, Tucker et al. 2018].

Outro efeito similar mas muito menos estudado é o conceito de politização, fenômeno que frequentemente acompanha a polarização e que se refere ao ato de fazer algo se tornar político [Wiesner 2021], sendo assim uma outra dimensão que

molda os comportamentos das pessoas. Temas recentemente sujeitos a uma crescente politização incluem as mudanças climáticas [Pepermans and Maeseele 2016], a COVID-19 [Hart et al. 2020], a religião [Zembylas et al. 2019] e a cultura e ciência em geral [Wright 1998, Bolsen and Druckman 2015]. Ao adicionar uma carga ideológica a uma questão não política, a politização pode levar à manipulação, aumento da hostilidade e falta de confiança no debate público.

Como a maioria dos estudos sobre discussões políticas focam em espaços já políticos [Rajadesingan et al. 2021, Wojcieszak and Mutz 2009], se perde a oportunidade de se analisar a politização. Por isso, nesse projeto, são estudados contextos políticos e não políticos de três redes sociais: TikTok, Twitter e YouTube. Isso permite a observação, de maneira mais ampla, do comportamento político, incluindo a observação da politização.

Como forma de medir e detectar politização, é utilizado um método baseado no conceito de *topic shift*, observado nas replies aos posts das diferentes redes sociais. Um *topic shift* ocorre quando em um determinado ponto de uma discussão, há uma mudança de tópico [Konigari et al. 2021]. Dessa forma, tópicos que apresentam grande probabilidade de *topic shift* em direção à tópicos políticos seriam mais politizados. Por meio da análise proposta, pretendemos definir se as redes sociais estudadas apresentam politização e como isso varia nas diferentes plataformas.

2. Dataset

Buscando compreender o comportamento de usuários brasileiros no contexto das eleições de 2022, nós coletamos dados de três plataformas diferentes: Twitter, YouTube, e TikTok. Embora as duas primeiras tenham sido alvo de vários de estudos ao longo dos anos, a última é uma plataforma que tem crescido rapidamente, e, como tal, o comportamento dela e de seus usuários ainda é pouco compreendido pela comunidade científica [Montag et al. 2021, Ling et al. 2022].

Nossos dados consistem em postagens (ou seus equivalentes em uma determinada plataforma) e as reações a elas (likes, replies). Esses posts foram coletados dos veículos de notícia com mais alcance nas redes sociais estudadas. Acreditamos que estes grupos nos permitem analisar conteúdos políticos, não políticos e ”quase políticos”.

3. Resultados e analises

Com base nas características de politização discutidas anteriormente, os datasets das redes sociais foram analisados visando responder três questões.

1. Com qual frequência ocorre a politização no contexto das eleições brasileiras?
2. Quais tópicos estão mais sujeitos à politização?
3. Há alguma diferença nas redes estudadas em relação a esse fenômeno?

As metodologias e resultados completos estão presentes no artigo desenvolvido para a conferência ICWSM 2024 como parte do projeto, cujo manuscrito estará anexado ao fim desse resumo.

Classificando textos de rede social: Para permitir a medição da frequência de politização baseada em *topic shifts* em direção à política, é necessário primeiro determinar o que pode ser considerado político. Para isso, foi treinado um classificador binário

que classifica comentários e vídeos em político(P) ou não político(NP). O treino de um classificador desse tipo de maneira tradicional utilizaria milhares de textos rotulados, os quais não eram acessíveis. Por isso, optou-se por um classificador baseado em Positive and Unlabeled(PU) Learning. Especificamente, utilizou-se um método de PU-learning em duas etapas [Bekker and Davis 2020].

Nesse método, é necessário se iniciar com um conjunto de dados P, normalmente manualmente anotados. Porém, devido à característica do domínio, observa-se que algumas palavras levam à comentários e posts quase sempre políticos, indicando que uma filtragem por keywords poderia ser suficiente para gerar esse conjunto inicial. As palavras-chave escolhidas para isso foram as palavras discriminativas mais frequentes na página das eleições de 2022 da Wikipédia[1], somadas com a hashtag política mais frequente no TikTok (#eleições2022).

Utilizando esse conjunto de dados e a técnica de PU learning em duas etapas baseada no conceito de *spies* apresentada em [Liu et al. 2002], foi possível treinar um classificador XGBoost [Chen and Guestrin 2016] com f1 de 0.82 considerando comentários e posts, que, embora não perfeito, permitiu a análise do fenômeno de interesse.

Politização nas diferentes redes: Após classificados todos os textos de todas as redes sociais, observa-se algumas tendências interessantes. Primeiramente, ao se comparar Twitter, YouTube e TikTok, é possível se notar que enquanto as duas primeiras redes apresentam uma proporção de notícias e comentários políticos parecida, a última rede apresenta consideravelmente menos conteúdo político, o que indica que os mesmos perfis de notícias postaram menos notícias políticas nessa rede. Isso pode indicar que há menos interesse em conteúdo político no TikTok em geral quando comparado com as demais redes.

Olhando para a proporção de comentários políticos em cada tipo de video, nota-se que novamente, no TikTok, notícias não-políticas tendem a ter uma menor proporção de comentários políticos do que nas demais redes. Essa diferença observada do TikTok em relação ao YouTube e Twitter pode ser explicada por alguns fatores, como por exemplo a natureza de vídeos mais curtos da plataforma que pode ser um deterrente de discussões mais profundas e a audiência mais nova do TikTok, que pode ter menos interesse em tópicos políticos. É importante notar que essas são apenas hipóteses e o real motivo para essas diferenças pode não ser nenhum desses.

Politização de tópicos: Utilizando BERTopics [Grootendorst 2022], que é um método de modelagem de tópicos baseado na arquitetura de transformers pré-treinada BERT para a produção de tópicos interpretáveis, foi possível definir quais temas levaram à uma maior quantidade de mudanças de tópico durante o período estudado.

Dentre os tópicos não-políticos menos politizados, destacam-se notícias relacionadas à entretenimento, esportes e famosos, como por exemplo NFL, futebol e notícias do Pelé, sendo que tópicos dessa natureza tiveram uma porcentagem $< 20\%$ de comentários políticos. Por outro lado tópicos mais politizados, no geral, incluíram temas mais controversos, como por exemplo educação, combustível, drogas e economia, tópicos dessa natureza em geral tiveram porcentagem de comentários políticos $> 45\%$.

[1] https://pt.wikipedia.org/wiki/Eleição_presidencial_no_Brasil_em_2022

Por outro lado, olhando para os tópicos políticos, nota-se que todos eles tiveram probabilidade de comentários políticos $> 50\%$, indicando que embora o classificador PU não seja perfeito, ele parece ser capaz de classificar notícias e comentários políticos corretamente na média. Dentre os tópicos políticos nota-se várias das ocorrências importantes no calendário eleitoral brasileiro, como por exemplo campanhas eleitorais, bloqueios nas rodovias e gastos governamentais.

Conclusão: A partir dessas análises, pode-se responder as perguntas de pesquisa. Destaca-se que a politização ocorre de maneira frequente no contexto das eleições brasileiras, com mais de 70% dos vídeos contendo pelo menos 1 comentário classificado como político e, alguns tópicos mais controversos chegando a ter 50% ou mais de comentários políticos. Também nota-se que a politização não ocorre de maneira igual em todas as redes sociais, sendo mais frequente no YouTube e Twitter.

Referências

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Mach. Learn.*, 109(4):719–760.

Bolsen, T. and Druckman, J. N. (2015). Counteracting the Politicization of Science. *Journal of Communication*, 65(5):745–769.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hart, P. S., Chinn, S., and Soroka, S. (2020). Politicization and polarization in covid-19 news coverage. *Science Communication*, 42(5):679–697.

Konigari, R., Ramola, S., Alluri, V. V., and Shrivastava, M. (2021). Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, Singapore and Online. Association for Computational Linguistics.

Layton, M. L., Smith, A. E., Moseley, M. W., and Cohen, M. J. (2021). Demographic polarization and the rise of the far right: Brazil's 2018 presidential election. *Research & Politics*, 8(1):2053168021990204.

Ling, C., Blackburn, J., De Cristofaro, E., and Stringhini, G. (2022). Slapping cats, bopping heads, and oreo shakes: Understanding indicators of virality in tiktok short videos. In *14th ACM Web Science Conference 2022*, pages 164–173.

Liu, B., Lee, W. S., Yu, P. S., and Li, X. (2002). Partially supervised classification of text documents. In *ICML*, volume 2, pages 387–394. Sydney, NSW.

Montag, C., Yang, H., and Elhai, J. D. (2021). On the psychology of tiktok use: A first glimpse from empirical findings. *Frontiers in public health*, 9:641673.

Pepermans, Y. and Maeseele, P. (2016). The politicization of climate change: problem or solution? *WIREs Climate Change*, 7(4):478–485.

Rajadesingan, A., Budak, C., and Resnick, P. (2021). Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, volume 15.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review*, 34(3):150–160.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Wiesner, C. (2021). *Rethinking Politicisation in Politics, Sociology and International Relations*. Palgrave Studies in European Political Sociology. Springer International Publishing.

Wojcieszak, M. E. and Mutz, D. C. (2009). Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 59(1):40–56.

Wright, S. (1998). The politicization of 'culture'. *Anthropology Today*, 14:7.

Zembylas, M., Loukaidis, L., and Antoniou, M. (2019). The politicisation and securitisation of religious education in greek–cypriot schools. *European Educational Research Journal*, 18(1):69–84.

Topic Shifts as a Proxy for Assessing Politicization in Social Media

Marcelo Sartori Locatelli,[1] Pedro Calais,[1] Matheus Prado Miranda*,[1] João Pedro Junho*,[1] Tomas Lacerda Muniz,[1] Wagner Meira Jr.,[1] Virgilio Almeida[1]

[1]Universidade Federal de Minas Gerais
6627 Av. Pres. Antônio Carlos
Belo Horizonte, Minas Gerais 31270-901 Brazil
publications23@aaai.org

Abstract

Politicization is a social phenomenon studied by political science, characterized by the extent to which ideas and facts are given a political tone. A range of topics, such as climate change, religion and vaccines has been subject to increasing politicization over the last few years, and social media platforms are a natural environment that enables the study of politicization as a core political process. In this work, we propose a computational method for assessing politicization in online conversations based on *topic shifts*, i.e., the degree to which people switch topics in online conversations. The intuition is that topic shifts from a non-political topic to politics are a *direct* measure of politicization – making something political, and that the more people switch conversations to politics, the more they perceive politics as playing a vital role in their daily lives. A fundamental challenge that must be addressed when one studies politicization in social media is that, a priori, *any* topic may be politicized. Hence, any keyword-based method or even machine learning approaches that rely on topic labels to classify topics are expensive to run and potentially ineffective. Instead, we learn from a seed of political keywords and use Positive-Unlabeled (PU) Learning to detect political comments in reaction to non-political news articles posted on Twitter, YouTube, and TikTok during the 2022 Brazilian presidential elections. Our findings indicate that all platforms show evidence of politicization as discussion around topics adjacent to politics such as economy, media behavior, education and drugs tend to shift to politics.

Introduction

Nowadays, any person may publicly share their views on a given subject with a far larger reach than they would have otherwise (Boynton and Richardson Jr 2016), social media platforms have enabled a plethora of studies in the social sciences and, more specifically, in the political sciences (Edelman et al. 2020; Lazer et al. 2009). While threats to the validity of studies based on social media data are still a concern (Howison, Crowston, and Wiggins 2011), access to large amounts of digital behavioral data has allowed political scientists to pair with their computer science peers to study the role of social media in government behavior (Graham, Avery, and Park 2015), voter engagement (Grover et al. 2019), news coverage and its bias (Oschatz, Stier, and Maier 2022; Baum and Groeling 2008) and even election forecasts (Tumasjan et al. 2011). More specifically, two widely recognized political processes received special attention with respect to how they shape (and are shaped) by social media, namely, *polarization* and *politicization*.

Polarization and politicization are two related but different political processes that directly impact how individuals allow their motivations and emotions to affect how they interpret new information (Druckman, Peterson, and Slothuus 2013; Taber, Cann, and Kucsova 2009). *Polarization* refers to the process by which two or more political groups, by selectively choosing to consume opinions they already agree with, adopt increasingly extreme and antagonistic viewpoints, creating the so-called echo chambers (Spohr 2017; Layton et al. 2021; Tucker et al. 2018).

Politicization, on the other hand, is the act of marking or naming something as political (Wiesner 2021). Topics recently subject to increasing politicization include climate change (Pepermans and Maeseele 2016), COVID-19 (Hart, Chinn, and Soroka 2020), religion (Zembylas, Loukaidis, and Antoniou 2019), and culture and science in general (Wright 1998; Bolsen and Druckman 2015). By adding an ideological charge to a non-political issue, politicization may lead to manipulation, increased hostility and a lack of trust to the public debate. Next we present a concrete example of politicization in a news article published by Folha de São Paulo newspaper in Twitter during the Brazilian 2022 Presidential Elections:

> **Post**: "Latin America: Evangelical gays defy churches and get married after pro-LGBTQIA+ referendum in Cuba."[1]
>
> **Comment**: "Recently, @folha's headlines have seemed to be made specifically to be used by Bolsonaro's supporters and fuel disinformation."[1]

Note that the original post touches on an (apolitical) religious and LGBTQIA+ topic that was quickly labeled as politically motivated – in particular, meant to be politically exploited by supporters of Brazilian 2022 presidential candidate Jair Bolsonaro. In this work, we devise a computational

*These authors contributed equally.

[1]Posts have been translated from Portuguese to English.The comment was paraphrased to protect the identity of the user.

method that directly models two key aspects that characterize politicization:

1. The transition from a non-political to a political topic. On unfiltered social media datasets comprising conversations on several topics, the post starting the discussion and the comments written in reaction to it can be both political and non-political; to detect politicization, we find *topic shifts* in which the original post is non-political, but comments are political, as a proxy for politicization.
2. The fact that, a priori, *any* non-political topic may be politicized, and, hence, we cannot predict or anticipate all topics that may become political; manual labeling of individual posts that cover the wide spectrum of non-political topics would be costly. We have a set of high-precision positive labeled posts and comments derived from unambiguous political keywords, but we do not have negative (non-political) labels. To address this challenge, we resort to a semi-supervised machine learning strategy that learns from positive and unlabeled examples known as Positive-Unlabeled Learning (PU learning for short) (Bekker and Davis 2020). The key capability of PU learning is that it works in the absence of negative training examples and finds a boundary between positive and (hidden) negative examples under the assumption that their feature distribution is different.

Our method extends existing strategies to study politicization in online media, which typically share two limitations: they are focused on a single, specific topic and are fully keyword-based or require negative (non-political) labels, which limits the extent to which topic shifts can be observed. By starting with a small seed of high-precision political keywords, but expanding them through a two-step PU Learning strategy, we were able to perform a general and broad characterization of politicization on social media which found, based on Twitter, YouTube, and TikTok data collected during the 2022 Brazilian Presidential Elections, that:

- By starting with a small seed of high-precision political keywords and using word2vec features in an XGBoost classifier, it is possible to reach a 82% F1 score to distinguish between political and non-political news posts and comments;
- Politicization is a widespread phenomenon on social media. While political content is less prevalent on TikTok, in all three platforms, at least one out of two non-political news posts will receive at least one political comment.
- Topics that are more heavily politicized include hard news such as the economy, media behavior, education, and drugs, but even soft news such as sports are politicized.

Our paper is organized as follows. Initially, we discuss related work on polarization and politicization and provide more details on how our research extends the existing literature. Next, we detail the datasets we use and the computational method we employ to find news articles that are highly politicized. Next, we use the model to characterize politicization along several dimensions, such as prevalence, topics, and time. Finally, we discuss conclusions and future research directions.

Related Work

Political science has deeply studied polarization and politicization through online social media data. We will briefly review the computational methods typically employed to measure and observe those two core political processes.

Polarization. Several observational studies of political polarization, i.e., the division of a group of people into two sharply contrasting sets of opinions or beliefs, have been conducted over the last years on platforms such as Twitter (Conover et al. 2011), YouTube (Bessi et al. 2016), Facebook (Del Vicario et al. 2016) and Instagram (Fernandes et al. 2020). Computational methods include either measuring if each political side consumes and shares different content (Garimella and Weber 2017; Grover et al. 2019), or employing network analysis to find clear and dense communities of users or content representing the opposing poles (Kubin and von Sikorski 2021; Calais et al. 2013; Chin, Coimbra Vieira, and Kim 2022).

Politicization. Observational studies that investigate the degree of politicization in society usually focus on a single non-political topic, which can be as specific as the adoption of a low-carb diet in Sweden (Holmberg 2015), a Star Wars movie (Bay 2018), or a mega sports event such as the World Cup (Meier et al. 2021); typical conclusions are that those topics have been subject to increasing politicization.

One can see if a non-political topic is made political by correlating it with polarization (Peterson and Muñoz 2022; Brummette et al. 2018; Weber, Garimella, and Borra 2013): if distinct political groups refer to a non-political topic differently or at different rates, it is a strong signal of a politicized topic, such as "gun violence" and "religious freedom", receiving different attention from Democrats and Republicans (Kane and Luo 2018). Another common strategy to evaluate politicization is to count the extent to which a piece of content mentions political actors (Chinn, Hart, and Soroka 2020; Hart, Chinn, and Soroka 2020); the more a non-political content (such as COVID-19) is linked to politician names or political concepts, the more politicized it is.

Our work. We extend and generalize existing politicization studies over social media data in two important directions:

1. Instead of focusing on a specific and delimited topic such as COVID (Diaz et al. 2022), we enable the study of politicization in general social media data that comprises *both* political and non-political news articles and look at comments that can also be political or non-political. Hence, we are able to assess the general prevalence of politicization in online spaces;
2. Current research focuses on inferring politicization by looking at whether a piece of political content is *mixed* into non-political content, e.g., if a news article about COVID mentions a political actor or if a political content is posted in a non-political community (Wojcieszak and Mutz 2009; Rajadesingan, Budak, and Resnick 2021). We look at a stronger, more explicit phenomenon: the

	# News Source Profiles	# Posts	# Avg. Comments per Post	# Avg. Comments per User	Data Collection Period
TikTok	41	8,814	20.28	1.37	2022-08-24 to 2022-11-01
Twitter	50	119,691	27.75	5.66	2022-08-26 to 2023-03-03
YouTube	43	12,616	347.80	4.61	2022-01-01 to 2023-05-06

Table 1: Statistics per social media platform. Based on profiles of relevant news sources in Brazil, we collected their posts and the comments posted by other users in reaction to each post.

actual *shift* of a previous non-political content to a political one, as illustrated by the shift of LGBTQIA+ topic to Politics in the Introduction.

To enable the observation of the politicization of any topic, we employ a classification approach fed with high-precision keywords related to Politics. Since we do not have access to labeled non-political content, we employ a semi-supervised approach called Positive-Unlabeled (PU) Learning. PU Learning has been used to learn from social media posts in the context of tasks such as sentiment analysis (Wang, Zhang, and Liu 2017), classification of user profiles (Karimi et al. 2021), and fake news detection (Liu and Wu 2020). Here, we use the technique to learn to classify unlabeled posts and comments, which can be either political or non-political.

Finally, after we build a classifier that distinguishes between political and non-political content, we look for *topic shifts*, i.e., comments that change the post's original topic to another related, but different topic. More specifically, we are interested in shifts from non-political topics to a political one, as this is a strong indicator of politicization. Topic shifts have been investigated from a linguistic perspective (Sun and Loparo 2019) and also with respect to their potential to bring frustration to discussion forums (Park et al. 2016). To the best of our knowledge, we are the first to explicitly connect topic shifts with the concept of politicization from a computational social science perspective.

Dataset: YouTube, Twitter and TikTok

Motivated by the highly polarized 2018 Brazilian Presidential Elections (Layton et al. 2021; Fernandes et al. 2020) and the extensive use of social media platforms by the presidential candidates, we study the politicization of news posts during the 2022 Elections. We collected data from three platforms: YouTube, Twitter, and TikTok. While the first two have been the target of a plethora of studies over the years (Montag, Yang, and Elhai 2021; Ling et al. 2022), the latter is a platform that has grown extremely fast recently, and, as such, the behavior of its users is still poorly understood by the research community (Medina Serrano, Papakyriakopoulos, and Hegelich 2020).

We collected all posts (and associated comments) published by popular Brazilian news sources (or their equivalents on a given platform) and the reactions to them (likes, shares, replies, and comments). Therefore, we observe not only political but also non-political news and associated comments, which enables a range of new perspectives on political behavior, including observing politicization, a concept that by its nature requires non-political data to be appropriately observed. To allow for an appropriate analysis, only comments with more than 5 tokens were considered.

The selection of news source profiles was conducted in order to select some of Brazil's most prominent digital news sources[2]. Therefore, the profiles with the most significant engagement, measured in followers or likes, were those picked for the research. Note that not all of them have profiles on each of the three platforms.

On YouTube and Twitter, these were collected using the official APIs, YouTube Data API v3 and Twitter API v2, respectively, made available by the platforms. On TikTok, due to the lack of an official API at the time of the collection period, an unofficial API[3] as well as web scrapers were used. Table 1 shows the statistics for the collected data. We aimed to collect news on all platforms well into 2023. However, due to TikTok updates in November 2022, the unofficial API and other ways of collecting data stopped working, interrupting the collection on that platform.

Detecting politicization with Positive-Unlabeled Learning

Manually inspecting or labeling posts in the datasets searching for political content and politicization of non-political content would be costly and time consuming. Differently from (Rajadesingan, Budak, and Resnick 2021), which focus on Reddit communities, we do not have social groups we could label as political or non-political, since, on YouTube, Twitter, and TikTok, the discussion is centered around individual content and not communities. However, in the context of (Brazilian) politics, it is fairly easy to identify some posts that are very unambiguously political, such as a post that cites Lula or Bolsonaro (the front-runners of the 2022 presidential elections). Additionally, in our dataset, especially on TikTok, *#eleicoes2022* was one of the most prevalent hashtags, referring to the presidential elections. Given these assumptions, by using the 10 most frequent words related to politics in the 2022 Brazilian general election Wikipedia page[4], as well as the aforementioned hashtag, it was possible to identify a positive set P composed of news and comments very strongly linked to Politics[5].

By using those high-precision political keywords, we can conduct a first examination of the prevalence of political news posts and associated comments. In Tables 2 and 3, we see that at least 13%, 22%, and 24% of news posts on Tik-

[2]See Appendix for the full list of profiles.

[3]https://github.com/davidteather/tiktok-api

[4]https://pt.wikipedia.org/wiki/Eleição_presidencial_no_Brasil_em_2022

[5]See the Appendix for the full list of political keywords.

Tok, Twitter, and YouTube are political, respectively. Comments posted in reaction to political posts tend to have between 2 and 3 times more political content than comments posted in reaction to unlabeled content, which is consistent with the expectation that political news attract more political comments. The interesting numbers, however, are those in Table 3: among unlabeled news posts, between 7 and 11% of the comments are political, and from 28 to 58% of comment threads contain at least one post that is unambiguously political, which may hide the politicization of non-political news posts.

		Comments in Political News Posts		
Platform	**Political News Posts**	Political	Unlabeled	At least one Political Comment
YouTube	24%	31%	69%	94%
Twitter	22%	15%	85%	76%
TikTok	13%	33%	67%	82%

Table 2: Ratio of political posts per platform and prevalence of political comments. Since we considered high-precision political keywords, these are approximate lower bounds for the prevalence of politics in the dataset.

		Comments in Unlabeled News Posts		
Platform	**Unlabeled News Posts**	Political	Unlabeled	At least one Political Comment
YouTube	76%	11%	89%	58%
Twitter	78%	7%	93%	33%
TikTok	87%	9%	91%	28%

Table 3: Ratio of unlabeled posts per platform and prevalence of political comments. Unlabeled posts can be either political or non-political.

Two-step PU learning

To actually assess politicization, we must be able to classify unlabeled content as political or non-political to some degree. Given a set P of positive examples about Politics and a set U of unlabeled examples (which contain hidden examples about Politics and content that is non-political), we want to build a classifier using P and U that can identify positive (political) and negative (non-political) documents in U (Liu 2007).

To operate in this semi-supervised setting, we employ a Positive-Unlabeled (PU) Learning strategy called *two-step* (Bekker and Davis 2020). In this strategy, we first (1) find reliable negative examples (non-political examples) and then (2) use supervised or semi-supervised techniques with the labeled, reliable negatives and, optionally, unlabeled examples as inputs. The underlying assumption here is that unlabeled positive examples are similar to their labeled counterparts, while negative examples are sampled from a different distribution.

First PU Learning step: extracting reliable negative examples with spies. To find reliable non-political (negative) examples, we use *spies* (Liu et al. 2002). Spies are a random selection of a fraction of the positive labeled examples (we used s% = 10% of positive examples), which will be treated as unlabeled examples. Since we know they are actually positive examples, we will use the probability scores attributed to the spies by the classifier trained on P (with spies removed) and U to calibrate the label probability that delimits the boundary that separates positive from negative examples – see Step 1 depicted in Figure 1.

In an ideal world, we would classify the reliable negatives as the examples that were attributed probabilities lower than $\min P[c = P|s_1], P[c = P|s_2]...P[c = P|s_k]$, where s_k is the k-th spy and c is the predicted class, we call this threshold t. However, due to the existence of noise and outliers, some spies may have lower probability than most negative documents, so a noise level l is used to estimate t so that $l\%$ of the documents have probability lower than t. We used $l = 15\%$ following advice from (Liu et al. 2002) that any rate between 5 and 20% works well.

The pseudo-code for PU Learning Step 1 is shown in Algorithm 1. We used TF-IDF as features and a Naive Bayes Classifier.

Algorithm 1: PU Learning Step 1 algorithm

$N \leftarrow \emptyset$ {Initialize Reliable Negatives}
$S \leftarrow$ sample($P, s\%$) {Initialize Spies}
$US \leftarrow U \bigcup S$
$P \leftarrow P - S$
Train Naive Bayes using US and P
Classify each document in US
Estimate t using S
for u_i in U **do**
  if $P[c = P|u_i] < t$ **then**
    $N \leftarrow N \bigcup \{u_i\}$
    $U \leftarrow U - \{u_i\}$
  end if
end for

Second PU Learning step: Traditional binary classifier. In the second step, we learn a traditional classifier fed with positive and the negative examples obtained from step 1 – see Figure 1. During our experiments, we tested a variety of word representations and classifiers, including a fine-tuned BERT model, however, our discussions will focus on word2vec as the word representation and gradient boosting as the classifier, using the XGBoost library (Chen and Guestrin 2016).

Baselines. To appropriately factor in the impact of the 2-step PU Learning solution for our political classification problem, we compare it with a few baselines:

1. A keyword-based classifier based on the political keywords to expose the extent to which a simple match of keywords is enough to separate political from non-political content;
2. A gradient boosted tree classifier that considers all unlabeled content as negative, to assess the actual need for treating unlabeled examples in a PU fashion;
3. A PU Learning strategy based on the incorporation of class priors to calibrate the classification[6] (Elkan and Noto 2008). This strategy, in principle, should not be adequate for our problem since the high-precision keywords

[6]Based on this implementation: github.com/pulearn/pulearn.

used to create P tend not to be a random sample of the full set of positive examples but rather a biased and easier-to-classify sample.

Experiments

To evaluate the PU Learning strategy to separate political from non-political content, we manually annotated 3,982 news posts and an equal number of comments, of which 1,744 were posted to Twitter, 1,225 to YouTube, and 959 to TikTok[7]. Of these, 2,194 news are political and 1,734 are not, while 2,210 comments are political and 1,718 are not.

We compare the performance of the baseline models and the PU variants: the political keyword classifier, an XGBoost that naively treats unlabeled examples as negative, and three flavors of PU Learning: one based on using class priors and two using the 2-step strategy, one using a fine-tuned PT-BR BERT model (Souza, Nogueira, and Lotufo 2020), and the other using XGBoost. Hyper-parameters were tuned using grid search. Accuracy, weighted average, recall, precision, and F1 for all models are summarized in Table 4, and we show the metrics separately for news posts and comments in Figures 2a and 2b, respectively.

The 2-step PU method with XGBoost is the best performing model. F1 score is 0.82, which is aligned with similar works that involved the use of two-step PU learning techniques in other contexts involving text (Fusilier et al. 2015; Li et al. 2014). Interestingly, all models perform better for the news posts (0.88 F1) when compared with comments (0.77 F1), which is expected as their text is not only longer, on average, than comments, but also has more context and structure.

Characterizing Politicization

Having built a classifier that separates news posts and comments into political and non-political, we characterize politicization by seeking out topic shifts. More specifically, we search for comments that were classified as political despite referring to a piece of non-political news.

Tables 5 and 6 show the prevalence of political comments for news predicted as political and non-political, respectively. Note in Table 5 that 94%+ of all news posts triggered at least one political comment, what is consistent to what we expect. Also, the proportion of political comments in response to political news increased from 31, 15, and 33% to 79, 68 and 72% for YouTube, Twitter and TikTok when we compare with Table 2,indicating that the classifier is indeed expanding the boundary of what political content looks like.

TikTok appears to be much less politicized, with a higher percentage of non-political news, coupled with a low percentage of political comments on those news, while also having the lowest percentage of news with at least one political comment among all platforms. It is worth noting that YouTube and Twitter appear to be much more similar to one another than TikTok, as the same news sources posted much more political news to the first two social media platforms while prioritizing non-political posts on the latter.

[7]Three of the authors independently produced labels, and we ran a majority vote.

The aforementioned differences and similarities between the studied platforms make it important to contrast the characteristics of topic shift on each of them, as their distinct features and public may influence this aspect. Figures 3a and 3b show the Cumulative Distribution Function of Topic Shifts on YouTube and TikTok comment sections, respectively. The distribution for Twitter was omitted due to it being very similar to the YouTube one.

These distributions show that topic shift is a phenomenon that is more prevalent in non-political news, leading the conversation to political topics. In fact, on both YouTube and Twitter, the non-political news CDF crosses the political news CDF on topic shift$\approx$20% and probability$\approx$50%. This means that the 50% of posts with the most topic shift in non-political news, show more topic shift than their political counterparts. TikTok, on the other hand, does not follow this pattern, with non-political and political news having similar topic shift distributions. This variation on TikTok could be explained by a number of factors, including, but not limited to:

- TikTok's short video format, which may inhibit deep or serious discussions;
- TikTok's younger audience (Kanthawala et al. 2022), which may be less interested in politics.

Temporal changes in Topic Shifts

Exploring the temporal dynamics of topic shifts is also important for better understanding how this metric relates to societal politicization. This section focuses on YouTube because it is the platform for which we have a long data collection period. Figure 4 shows the frequency of topic shifts in each week's comments. Some significant events concerning the Brazilian election are also highlighted.

We see a peak in topic shifts towards Politics roughly a week after the elections' result was announced and between the first and second election rounds, possibly due to the discussions between each candidate's supporters. Another even bigger peak is seen in the weeks prior to Lula taking office, which may be attributed to the political unrest surrounding the rumors of a coup. Politicization starts to reduce after Lula's inauguration as president, showing a small peak during the invasion of Congress.

The inverse tendency is seen in political news since, when the proportion of topic shifts is peaking on non-political news, it is trending downwards on political content. This may suggest that when the topic of politics is not in the spotlight, topic shifts become more frequent in political news, and when it is gaining traction, they become more prevalent in non-political news instead.

Finding Most Politicized Topics

The classifier described in the previous sections is able to predict when a piece of news or comment is political. This is enough to ascertain whether politicization happens or not in the context of online Brazilian news' comment sections. Now, we seek to find out which topics are more subject to

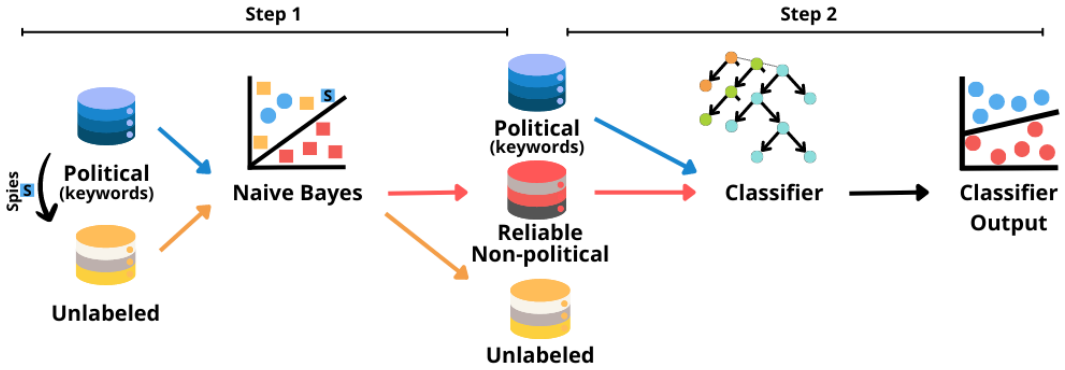


Figure 1: The two-step Positive-Unlabeled (PU) learning technique. Step 1 is fed with political and unlabeled examples and divides the unlabeled set into two sets – reliable non-political and a smaller unlabeled set. Step 2 is a traditional binary classifier fed with political and reliable non-political examples. Squares represent examples treated as unlabeled during the first step, while circles represent examples treated as labeled. Red represents examples classified as non-political, blue represents political and yellow, unlabeled.

Metric	Political Keywords	XGBoost with unlabeled	Class Prior XGBoost	Two-step BERT	Two-step XGBoost
Accuracy	0.70	0.70	0.71	0.75	**0.82**
F1	0.68	0.68	0.69	0.74	**0.82**
Recall	0.70	0.70	0.71	0.75	**0.82**
Precision	0.79	0.80	0.80	**0.83**	0.82

Table 4: Average score for each model considering news and comment predictions. Values in bold represent the best performing model for a given metric.

		Comments in Political News Posts		
Platform	**Political News Posts**	Political	Non-political	At least one Political
YouTube	71%	79%	21%	97%
Twitter	73%	68%	32%	94%
TikTok	35%	72%	28%	94%

Table 5: Ratio of political posts per platform predicted by the PU Learning-based classifier.

		Comments in Non-political News Posts		
Platform	**Non-political News Posts**	Political	Non-political	At least one Political
YouTube	29%	22%	78%	78%
Twitter	27%	37%	63%	63%
TikTok	65%	25%	75%	62%

Table 6: Ratio of non-political posts per platform predicted by the PU Learning-based classifier.

politicization. We identify topics using BERTopic (Grootendorst 2022), a topic modeling technique based on BERT embeddings and c-TF-IDF that produces interpretable topics.

The news posts were split into political or non-political based on the classifier output, and topics were produced for each of those two groups, with each topic covering at least 100 news posts. After assigning each news post to a topic, we calculate the percentage of comments for each topic whose classification was different from the content it referred to. We can then identify which topics were more likely to be shifted towards or away from politics, and since the topics are highly interpretable, we can manually spot misclassified news and exclude them from the analysis.

When looking at the topics generated by BERTopic, we can identify a variety of relevant events that happened in 2022/2023. On non-political news, we identify 48 topics, with examples such as soccer, cryptocurrencies, the NFL, and even Gisele Bündchen and Tom Brady's divorce. Meanwhile, on political news, we identify 28 topics, including elections, corruption (in a variety of areas), and candidate debates.

Before discussing the most politicized topics, it is important to acknowledge some possibly misclassified topics. On the non-political topics, we see protests in Iran, North Korean missiles, accounts blocked in social media due to political statements, trucker road blockades (when the focus is on the events and not politics), the war in Ukraine, and daily news (which most of the time contain political segments that may lead to political comments).

That said, excluding these possibly misclassified topics, the most politicized topics are the economy, disasters, media, drugs, and education. Meanwhile, the least politicized topics related to entertainment and lifestyle, with topics such as sports, pregnancy, food, and celebrities. Tables 7 and 8 show, respectively, the least and most politicized non-political topics. It is possible to see an extreme difference in the percentage of political comments, with the least politicized having less than 20% of comments classified as political, while the most politicized have over 50% political

Topic	Representative Words	Political Comments	Sample
NBA	Basketball, playoffs, Lakers	5%	**News.**"Video shows Draymond Green punching Jordan Poole during Golden State Warriors practice." **Politicized Comment.**"And how's the former prisioner?"(Lula)[1]
NFL	Quarterback, football, 49ers	7%	**News.**"MISSSSSSSS Matt Prater misses NFL's 50-yard field goal, Patriots vs. Cardinals still tied after Ari Aguiar's 'hex' #ESPNnoStarPlus" **Politicized Comment.**"We are suffering, Bolsonaro's help in Brasilia is missing."[1]
Soccer	Sports, goal, Atlético	10%	**News.**"PALMEIRAS 1 X 1 FLAMENGO | Best Moments | Brasileirão 2022 round 23." **Politicized Comment.**"Bolsonaro made Brasil worse."[1]
Space	Spaceship, planet, orbit	20%	**News.**"Science: Moon formation took place in a few hours, new NASA simulation suggests" **Politicized Comment.** "Lula supports censorship, tried to legalize abortions, persecute the church..."[1]
Pelé	Pelé, infection, diagnosis	20%	**News.**"Brazilian players pay tribute to Pelé after defeating Korea #FIFAWorldCup #Qatar2022" **Politicized Comment.**"Dance for Lula's victory...The homage to Pelé shows that despite political differences we are all Brazilian people"[1]

Table 7: Least politicized non-political topics across all platforms. Note how most of the topics relate to soft news.

comments. Notice that, although all topics show politicized comments, a fair amount of nitpicking was required to find politicization in Table 7, with many of the videos not having even a single political comment, while in Table 8 sampling a single random video and 2 or so comments classified as political was enough to find very explicit examples of politicization.

Comparing non-political with political topics, it can be seen that even the least politicized political topics have a percentage of political comments comparable to the most politicized non-political topics. In fact, only the protests in Iran, the Chinese government, and the war in Ukraine (focusing on politics) are less politicized than the most politicized non-political topic. All other topics have a greater percentage of political comments. This finding suggests that, while the classifier is not perfect, it is able to give higher probabilities in general for comments in political news. This is evidenced by the fact that non-political news have 26% of political comments, even when considering misclassified topics, while political news have 76% political comments.

Interestingly, some of the most politicized topics include religion in politics, elections, debates, and protests (focusing on politicians' reactions), which were very relevant topics in the context of the 2022 Brazilian elections, with candidates, such as Father Kelmon and many others, appealing to voters through religion. Moreover, some topics that previous studies defined as politicized were classified as such even when they did not explicitly use the keywords. Examples of this include vaccination movements and climate change, with some news that might seem innocuous, such as "COP-27: Why Greta Thunberg is avoiding the UN climate conference this year", being extremely politicized in the comments, in part due to Bolsonaro's previous clash with Greta.

Conclusions and Future Work

Politicization is the act of **transporting** an issue or an institution into the sphere of politics – making **previously unpolitical** matters political (Zürn 2019).

We study topic shifts over social media conversations as a novel strategy to measure politicization, more specifically in the context of the 2022 Brazilian presidential elections. While politicization is often studied on specific topics or mentioned in a cursory way, we propose a computational method that directly observes and quantifies politicization using *topic shifts*, i.e., the change of a topic by social media users participating in a discussion.

Starting from a few political keywords that work as seeds, we conduct a two-step PU Learning strategy that learns the boundary between political and non-political content. We evaluate the results against an annotated dataset, and our method achieves around 88% and 77% F1 scores on news posts and comments, respectively.

Our computational method enables the study of politicization of social media data comprising a set of arbitrary, previously unknown topics and our results indicate that, indeed, politicization is a prevalent social process in social media, aligned with previous research on Reddit communities (Rajadesingan, Budak, and Resnick 2021).

When looking at the politicization of news, our findings suggest that some topics are more politicized than others. For example, the economy, education, and drugs are much more politicized than topics related to entertainment and lifestyle, such as sports, pets, and food. However, every topic has at least some degree of politicization if you search for it, even if it is far from Politics.

We believe our work solidifies a recent trend that, since political talk may occur anywhere (Rajadesingan, Budak, and Resnick 2021), looking for behaviorial patterns when topics drift and merge gives us the opportunity to contrast behaviors, build null models, and compare the observed political behavior with that of control groups. For example, our results reinforce how motivated reasoning – the influence

Topic	Representative Words	Political Comments	Sample
Fossil Fuel	Prices, fuel, oil	60%	**News.**"How the decline in oil affects Petrobras, who plans to increase production of the fossil fuel." **Politicized Comment.** "If it is up to Paulo Guedes and Lula, Petrobrás will be sold.Only Ciro Gomes can save it."[1]
Economy	Inflation, economy, recession	54%	**News.**"Ceasa in Rio de Janeiro catches fire; warehouses are looted." **Politicized Comment.** "People are already blaming Lula and he is not even president yet."[1]
Media	News, reporter, press	51%	**News.**"Reporter suffered a pressure drop during a live link in São Vicente, on the coast of São Paulo #g1." **Politicized Comment.**"This would not happen in Jair's government."[1]
Education	Education, schools, learning	50%	**News.**"The pandemic led to setback in literacy in Brazil, MEC points out." **Politicized Comment.**"...congratulations to Bolsonaro who articulated corruption with evangelical priests..."[1]
Drugs	Drugs, cocaine, smuggling	50%	**News.**"PF arrests man with almost 1 ton of marijuana on Via Dutra, in Rio" **Politicized Comment.**"...Do people arrested for drug dealing vote for Bolsonaro?"[1]

Table 8: Most politicized non-political topics across all platforms, excluding misclassifications. Note how most of the topics relate to hard news.

of our motivations and goals in our reasoning – is a cognitive process that is highly tied to politicization (Bolsen and Druckman 2018).

In future work, we plan to better link politicization with polarization and attempt to establish potential correlations and cause-and-effect connections between those two core political processes. We will also be looking at user profiles; are there a few users that politicize everything?

Broader perspective, ethics and competing interests

All data we use from TikTok, Twitter, and YouTube was publicly available when we collected it. Additionally, all labels were created by people directly involved in the research project. To avoid compromising individual users, any comment quoted on this paper was translated, paraphrased, and modified (while keeping the general meaning).

Our work focuses on assessing and characterizing politicization without using any manual labels, which can accelerate and encourage further research in the political sciences. While we acknowledge that the accuracy of the classifier in the range of 82% is a potential threat to the validity of the results, since a correct topic shift is a result of a correct classification of both the news posts and the comment, we believe the effect of the prediction error is minimized due to two efforts: (1) we manually discarded the misclassified topical clusters of news posts, and (2) since each post receives on average tenths or hundreds of comments (Table 1), the errors tend to cancel out and a signal of politicization still emerges, as the analysis of topics in Tables 7 and 8 made clear.

References

Baum, M. A.; and Groeling, T. 2008. New Media and the Polarization of American Political Discourse. *Political Communication*, 25(4): 345–365.

Bay, M. 2018. Weaponizing the haters: The Last Jedi and the strategic politicization of pop culture through social media manipulation. *First Monday*.

Bekker, J.; and Davis, J. 2020. Learning from Positive and Unlabeled Data: A Survey. *Mach. Learn.*, 109(4): 719–760.

Bessi, A.; Zollo, F.; Del Vicario, M.; Puliga, M.; Scala, A.; Caldarelli, G.; Uzzi, B.; and Quattrociocchi, W. 2016. Users Polarization on Facebook and Youtube. *PLoS ONE*, 11.

Bolsen, T.; and Druckman, J. N. 2015. Counteracting the Politicization of Science. *Journal of Communication*, 65(5): 745–769.

Bolsen, T.; and Druckman, J. N. 2018. Do partisanship and politicization undermine the impact of a scientific consensus message about climate change? *Group Processes & Intergroup Relations*, 21(3): 389–402.

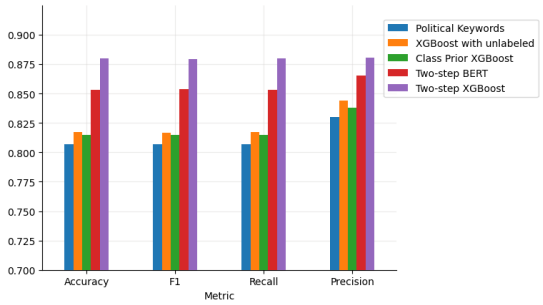
Boynton, G.; and Richardson Jr, G. W. 2016. Agenda setting in the twenty-first century. *New Media & Society*, 18(9): 1916–1934.

Brummette, J.; DiStaso, M.; Vafeiadis, M.; and Messner, M. 2018. Read all about it: The politicization of "fake news" on Twitter. *Journalism & Mass Communication Quarterly*, 95(2): 497–517.

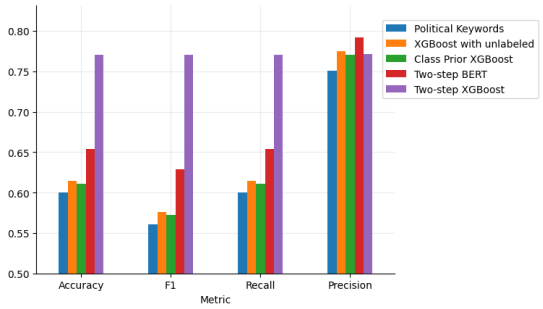
Calais, P.; Meira Jr, W.; Cardie, C.; and Kleinberg, R. 2013. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, volume 7, 215–224.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM. ISBN 978-1-4503-4232-2.

Chin, A.; Coimbra Vieira, C.; and Kim, J. 2022. Evaluating Digital Polarization in Multi-Party Systems: Evidence from

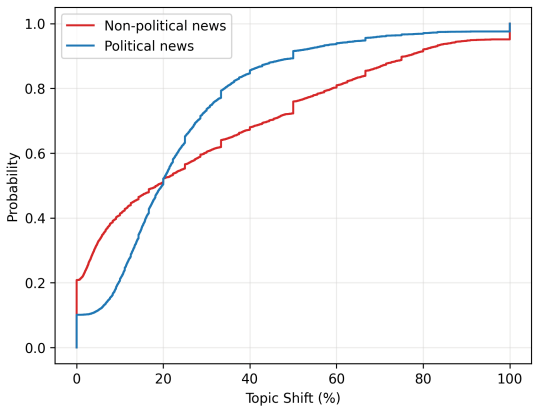


(a) News posts

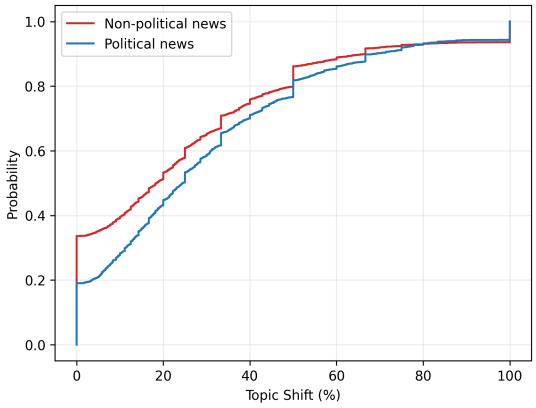


(b) Comments

Figure 2: Performance of the political classification models on news posts and comments, based on all three platforms. Performance is superior for news posts, possibly due to comments having less context and structure than well-formed news headlines.



(a) YouTube



(b) TikTok

Figure 3: Probability distribution of Topic Shifts on YouTube and TikTok comment sections per original video predicted class

the German Bundestag. In *14th ACM Web Science Conference 2022*, WebSci '22, 296–301. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391917.

Chinn, S.; Hart, P. S.; and Soroka, S. 2020. Politicization and Polarization in Climate Change News Content, 1985-2017. *Science Communication*, 42(1): 112–129.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, 89–96.

Del Vicario, M.; Vivaldo, G.; Bessi, A.; Zollo, F.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1): 37825.

Diaz, M. I.; Hanna, J. J.; Hughes, A. E.; Lehmann, C. U.; and Medford, R. J. 2022. The Politicization of Ivermectin Tweets During the COVID-19 Pandemic. *Open Forum Infectious Diseases*, 9(7). Ofac263.

Druckman, J. N.; Peterson, E.; and Slothuus, R. 2013. How Elite Partisan Polarization Affects Public Opinion Formation. *American Political Science Review*, 107(1): 57–79.

Edelman, A.; Wolff, T.; Montagne, D.; and Bail, C. A. 2020. Computational Social Science. *Annual Review of Sociology*, 46. 00000.

Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220.

Fernandes, C. M.; Ademir de Oliveira, L.; Motta de Campos, M.; and Gomes, V. B. 2020. Political polarization in the Brazilian Election Campaign for the Presidency of Brazil in 2018: an analysis of the social network Instagram. *Int'l J. Soc. Sci. Stud.*, 8: 119.

Fusilier, D. H.; Montes-y Gómez, M.; Rosso, P.; and Cabrera, R. G. 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information processing & management*, 51(4): 433–443.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, 528–531.

Graham, M. W.; Avery, E. J.; and Park, S. 2015. The role of social media in local government crisis communications. *Public Relations Review*, 41(3): 386–394.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

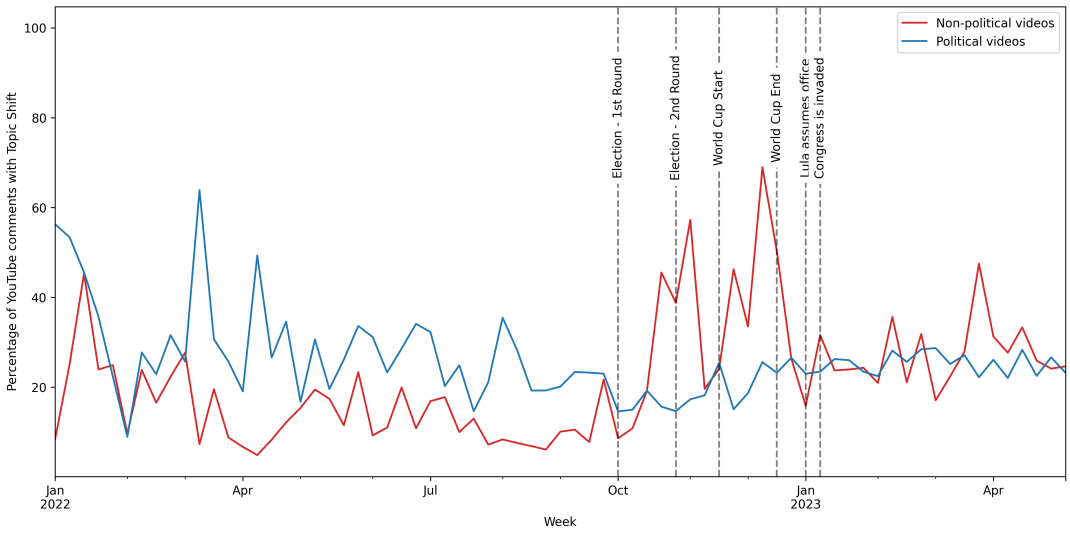


Figure 4: Percentage of YouTube comments which are a topic shift from the original news posts, by week. Between the Brazilian Elections 1st and 2nd rounds, we see a spike on non-political content being politicized. The FIFA World Cup correlates with another spike of politicization.

Grover, P.; Kar, A. K.; Dwivedi, Y. K.; and Janssen, M. 2019. Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145: 438–460.

Hart, P. S.; Chinn, S.; and Soroka, S. 2020. Politicization and Polarization in COVID-19 News Coverage. *Science Communication*, 42(5): 679–697.

Holmberg, C. 2015. Politicization of the Low-Carb High-Fat Diet in Sweden, Promoted On Social Media by Non-Conventional Experts. *Int. J. E-Polit.*, 6(3): 27–42.

Howison, J.; Crowston, K.; and Wiggins, A. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12.

Kane, B.; and Luo, J. 2018. Do the Communities We Choose Shape our Political Beliefs? A Study of the Politicization of Topics in Online Social Groups. In *2018 IEEE International Conference on Big Data (Big Data)*, 3665–3671.

Kanthawala, S.; Cotter, K.; Foyle, K.; and DeCook, J. R. 2022. It's the Methodology For Me: A Systematic Review of Early Approaches to Studying TikTok. In *HICSS*, 1–17.

Karimi, H.; Tang, J.; Weiss, X.; and Huang, J. 2021. Automatic Identification of Teachers in Social Media using Positive Unlabeled Learning. In *2021 IEEE International Conference on Big Data (Big Data)*, 643–652.

Kubin, E.; and von Sikorski, C. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3): 188–206.

Layton, M. L.; Smith, A. E.; Moseley, M. W.; and Cohen, M. J. 2021. Demographic polarization and the rise of the far right: Brazil's 2018 presidential election. *Research & Politics*, 8(1): 2053168021990204.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Alstyne, M. V. 2009. Computational Social Science. *Science*, 323(5915): 721–723.

Li, H.; Chen, Z.; Liu, B.; Wei, X.; and Shao, J. 2014. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE international conference on data mining*, 899–904. IEEE.

Ling, C.; Blackburn, J.; De Cristofaro, E.; and Stringhini, G. 2022. Slapping Cats, Bopping Heads, and Oreo Shakes: Understanding Indicators of Virality in TikTok Short Videos. In *14th ACM Web Science Conference 2022*, 164–173.

Liu, B. 2007. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer. ISBN 978-3-540-37882-2.

Liu, B.; Lee, W. S.; Yu, P. S.; and Li, X. 2002. Partially supervised classification of text documents. In *ICML*, volume 2, 387–394. Sydney, NSW.

Liu, Y.; and Wu, Y.-F. B. 2020. FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM Trans. Inf. Syst.*, 38(3).

Medina Serrano, J. C.; Papakyriakopoulos, O.; and Hegelich, S. 2020. Dancing to the partisan beat: A first analysis of political communication on TikTok. In *12th ACM conference on web science*, 257–266.

Meier, H. E.; Mutz, M.; Glathe, J.; Jetzke, M.; and Hölzen, M. 2021. Politicization of a contested mega event: The 2018 FIFA World Cup on Twitter. *Communication & Sport*, 9(5): 785–810.

Montag, C.; Yang, H.; and Elhai, J. D. 2021. On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in public health*, 9: 641673.

Oschatz, C.; Stier, S.; and Maier, J. 2022. Twitter in the News: An Analysis of Embedded Tweets in Political News Coverage. *Digital Journalism*, 10(9): 1526–1545.

Park, A.; Hartzler, A. L.; Huh, J.; Hsieh, G.; McDonald, D. W.; Pratt, W.; et al. 2016. "How did we get here?": topic drift in online health discussions. *Journal of medical Internet research*, 18(11): e6297.

Pepermans, Y.; and Maeseele, P. 2016. The politicization of climate change: problem or solution? *WIREs Climate Change*, 7(4): 478–485.

Peterson, E.; and Muñoz, M. 2022. "Stick to Sports": Evidence from Sports Media on the Origins and Consequences of Newly Politicized Attitudes. *Political Communication*, 39(4): 454–474.

Rajadesingan, A.; Budak, C.; and Resnick, P. 2021. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, volume 15.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Cerri, R.; and Prati, R. C., eds., *Intelligent Systems*, 403–417. Cham: Springer International Publishing. ISBN 978-3-030-61377-8.

Spohr, D. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review*, 34(3): 150–160.

Sun, Y.; and Loparo, K. 2019. Topic Shift Detection in Online Discussions using Structural Context. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, 948–949.

Taber, C. S.; Cann, D.; and Kucsova, S. 2009. The motivated processing of political arguments. *Political Behavior*, 31: 137–155.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2011. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Soc. Sci. Comput. Rev.*, 29(4): 402–418.

Wang, Y.; Zhang, Y.; and Liu, B. 2017. Sentiment lexicon expansion based on neural pu learning, double dictionary lookup, and polarity association. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Weber, I.; Garimella, V. R. K.; and Borra, E. 2013. Inferring audience partisanship for youtube videos. In *Proceedings of the 22nd International Conference on World Wide Web*, 43–44.

Wiesner, C. 2021. *Rethinking Politicisation in Politics, Sociology and International Relations*. Palgrave Studies in European Political Sociology. Springer International Publishing. ISBN 9783030545444.

Wojcieszak, M. E.; and Mutz, D. C. 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 59(1): 40–56.

Wright, S. 1998. The Politicization of 'Culture'. *Anthropology Today*, 14: 7.

Zembylas, M.; Loukaidis, L.; and Antoniou, M. 2019. The politicisation and securitisation of religious education in Greek–Cypriot schools. *European Educational Research Journal*, 18(1): 69–84.

Zürn, M. 2019. Politicization compared: at national, European, and global levels. *Journal of European Public Policy*, 26(7): 977–995.

Collected News Sources

Channel	Reach	Platform With Most Followers
g1	14.8 mi	Twitter
VEJA	9.1 mi	Twitter
Folha de S.Paulo	8.8 mi	Twitter
Estadão	7.5 mi	Twitter
Jornal O Globo	7.3 mi	Twitter
Jovem Pan News	7.3 mi	YouTube
ge	6.3 mi	Twitter
Globo News	5.6 mi	Twitter
UOL Notícias	5.2 mi	Twitter
ESPN Brasil	5.2 mi	Twitter
R7	5.1 mi	Twitter
CNN Brasil	4.0 mi	YouTube
Jornal da Record	3.9 mi	Youtube
Metrópoles	3.6 mi	TikTok
Pânico Jovem Pan	3.6 mi	YouTube
BBC News Brasil	3.4 mi	Twitter
UOL	3.4 mi	YouTube
Valor Econômico	2.6 mi	Twitter
Revista Oeste	1.3 mi	YouTube
GZH	1.1 mi	Twitter
Correio Braziliense	0.9 mi	Twitter
O TEMPO	0.5 mi	YouTube
Estado de Minas	0.5 mi	Twitter
A TARDE	0.5 mi	Twitter
SuperesportesMG	0.2 mi	Twitter

Political Keywords

- partido
- presidencia
- bolsonaro
- lula
- candidatura
- #eleicoes2022
- eleicoes
- eleitoral
- presidente
- debate
- eleicao