Mapping Anti-Vaccine Activism: Semantic Similarity to Access Disinformation Communities in Twitter (X) During the COVID-19 Pandemic in Brazil

Anonymous ACL submission

Abstract

This article uses the semantic similarity between fake tweets about COVID-19 vaccines in Portuguese to create a graph and identify disinformation communities on Twitter (currently X). All 2,857,908 tweets in Portuguese containing the word *vacina* (vaccine in Portuguese) were scrapped from October 30, 2020, to May 25, 2021. A BERT-based algorithm was used to identify fake tweets and obtain their cosine similarity. The study identified five main disinformation communities, highlighting central figures and their influence within these groups. Each community had a clear central subject, and four had a well-defined central spreader of disinformation. Seven of the ten most central users were banned from Twitter for violating community guidelines. This work shows that semantic similarity can be a powerful tool to map disinformation communities in social networks.

1 Introduction

Twitter (currently X) is a social network widely used during the COVID-19 pandemic (Cinelli et al., 2020). In particular, it was a very effective tool for spreading vaccine disinformation in Brazil (Ceron et al., 2021). Since the beginning of the pandemic, the Brazilian federal government undermined the severity of the disease to justify keeping the economy running (Ricard and Medeiros, 2020). This was only possible by denying scientific evidence for mitigation measures. Government officials, including the president, supported early treatment with ineffective drugs and did not mandate social distancing or the use of masks in public places. When vaccines were made available, the official recommendation was to challenge their safety and efficacy (Galhardi et al., 2020). The president himself claimed he would not be vaccinated. In a polarized society, this induced a clear correlation between political preference and antivax activism. In this context, false information concerning inefficacy and supposed severe side effects was intentionally spread to cause panic and vaccination hesitation among the population. There is evidence that false information diffuses faster than truth (Vosoughi et al., 2018). Without restrictions on disseminating false content, social networks became crucial to reducing the number of vaccinated persons and increasing active anti-vax activism. In the process, lay people genuinely looking for valuable scientific information may have been co-opted into disinformation bubbles. This is especially harmful because, at the time, social networks had surpassed traditional newspapers or television networks as the primary source of information in Brazil (Newman et al., 2024). Like many recommendation algorithms in social networks, the one used in Twitter[1] induces the formation of ideological bubbles. Users who like tweets on some emotional subject will be primarily exposed to tweets that match their ideological preferences and consequently interact only with people who think similarly (Ribeiro et al., 2018). We assume that a user emotionally touched by a text based on disinformation will somehow reproduce its ideas and form when writing his own tweets, providing positive feedback for forming isolated ideological bubbles. In this article, we used a criterion of semantic similarity between tweets to reveal the main features of the sub-network of fake tweets about vaccines in Portuguese during the COVID-19 pandemic. To the authors' knowledge, this is the first time that semantic similarity in texts is used to express the connection between users in a disinformation community.

2 Background

To map the COVID-19 vaccine disinformation community in Portuguese, a set of fake tweets about

[1] https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

vaccines (Geurgas and Tessler, 2024) was used. The dataset has been anonymized to contain no personal information. The dataset has offensive content, such as swear words and hate speech toward some ethnicities. Hate speech has not been removed or censored since this type of writing is often associated with fake content. No demographic or groups represented selection was used when collecting the data. The only criterion used was that the person must have a Twitter account at the time.

1. All 2,857,908 tweets in Portuguese containing the word *vacina*, vaccine in Portuguese, between October 30, 2020, and May 25, 2021 were scrapped using the old Twitter API[2]. This corresponds to a period of intense activity about vaccines in social networks in Brazil and the resurgence of the Brazilian anti-vax movement.

2. An automatic classifier of true/fake tweets was implemented by fine-tuning BERTimbau (Souza et al., 2020), a pre-trained BERT model for Brazilian Portuguese. This involved curating a random subset of 16,731 tweets. The classifier identified 932,666 fake tweets. Out of the fake Tweets, 264,378 were unique.

3 Methods

The main objective was to build a graph that represents the superstructure of users who posted fake tweets similar to tweets they had read.

3.1 Graph of fake tweets spreaders

The fake tweets community was represented by a graph in which the nodes were users and the edges represented the weight of the user interaction. The user interaction was expressed through an adjacency matrix. An adjacency matrix is a square matrix used in graph theory to represent a finite graph (Newman, 2018). For a graph with n vertices, the adjacency matrix is an $n \times n$ matrix where the element a_{ij} represents the weight w between vertices i and j. In an undirected graph, the matrix is symmetric, with $a_{ij} = a_{ji}$. When there is no connection between nodes i and j, $a_{ij} = 0$.

The Louvain method (Blondel et al., 2008) was used to reveal the community structure within the adjacency matrix. This algorithm detects communities even within large networks by maximizing an indicator of modularity, defined as the density of links inside communities compared to links between communities.

Once communities are detected, the importance of each of its members can be evaluated using metrics such as their degree of centrality, closeness centrality, and eigenvector centrality.

The degree of centrality of a node is defined as the number of edges connected to the node divided by the total number of edges on the graph (Newman, 2018). In social networks, it indicates the importance of nodes relative to their environment. In our study, the most central nodes are the primary sources of fake tweets that other nodes read and reproduce.

Closeness centrality measures how short the shortest path from node i to all other nodes (Entringer et al., 1976). Except when the nodes are only partially connected, it does not differ much from the degree of centrality.

The eigenvector centrality is proportional to the sum of the centralities of the node i neighbors (Newman, 2018). It measures the importance of a given node's first neighbors (the nodes that are directly connected). In a social network, it expresses the consolidated influence of a node with a high degree of centrality by evaluating whether the nodes influenced by it are also influent.

3.2 Weights

The graph weights w were determined by first calculating the inner product between tweets the users wrote. This was achieved by taking advantage of the vectorization of sentences by the BERT framework.

3.2.1 Vector Space

BERT produces word representations that are dynamically informed by the words around them. This characteristic of Transformers (Vaswani et al., 2017) allows us to estimate text similarity with a high confidence level by calculating their normalized inner product. The output dimensionality of the BERTimbau encoder stack is 768. Therefore, once embedded, the tweet output exists as vectors within an Euclidean $\mathbb{R}^{768}$ space (Reif et al., 2019).

Sentence-BERT (Reimers and Gurevych, 2019) was used to embed tweets in vectors within the BERTimbau base space. The internal products were obtained using the cosine similarity function

[2]The data collection occurred before Twitter (currently X) changed the API and limited free access for academic users.

defined by

$$\cos(\theta) = \frac{\vec{u}.\vec{v}}{\|\vec{u}\|\|\vec{v}\|}. \quad (1)$$

The cosine similarity function has a range of [-1,1]. A value of -1 indicates opposite vectors (representing semantically opposed tweets), while a value of 1 corresponds to identical vectors (representing the same tweets).

The output of the *SentenceTransformer*[3] function is a similarity matrix that consists of a square matrix with size input lines x columns. Each matrix element is the cosine similarity between the row i (tweet i) and column j (tweet j). Calculating the whole similarity matrix for large datasets is infeasible. The problem complexity is $O(n^2)$. Instead of computing all pairwise cosine similarities, the *paraphrase_mining* function of the *sentence_transformers* library can be used. This function divides the corpus into smaller chunks and returns only the pairs with higher cosine similarity. To avoid the trivial unity cosine distance between identical tweets, all tweets identical to tweets that had appeared before in the dataset were removed. This happens in two situations: when a tweet is actually a retweet (the user relays a previous tweet) or, more rarely, when a user reproduces a previous tweet verbatim. Discarding duplicate tweets also helps to reduce the occurrence of tweets sent by bots or by human users who relay tweets instead of actively interacting. To further minimize the influence of bots and irrelevant users, tweets from users with less than 20 followers or who have tweeted less than ten times during the data acquisition period were discarded. Some low-relevance legitimate users may have been lost in this context. The cosine similarities between tweets originating from the same user were also not computed, as they would be useless.

3.3 From Tweets to Users

To build a graph having users rather than tweets as nodes, a criterion to translate tweet cosine similarity to user link weight w must be chosen. Whenever there was more than one pair of similar tweets, the weight w was obtained by computing the average between the cosine distances of the different tweets.

[3] https://sbert.net/ Version 3.2.0.

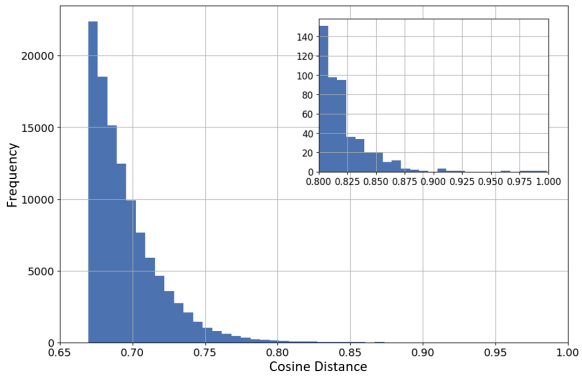

Figure 1: Distribution of cosine distances truncated at 0.67. The inset shows the details of the same distribution beginning at $minSim = 0.80$.

4 Results

The *paraphrase_mining* function used to calculate the cosine distance between tweets limits its results to values above a specific value that depends on both the query and corpus chunk sizes. With the parameters used in this work, only pairs with similarity above 0.67 were returned. This further reduced the database to 110,611 unique tweets.

The cosine similarity distribution is represented in Figure 1. The inset provides a closer view of the distribution above 0.80.

To visualize community formation, the window of similarities considered to generate the graph must be truncated at some value $minSim$. If $minSim$ is too low, the graph will be too dense, and less important communities will appear. If $minSim$ is too high, the graph is too sparse, and communities may not be very clear to isolate. We varied $minSim$ between 0.70 and 0.95 with increments of 0.05. Although the results do not depend critically on $minSim$ in this range, considering the computational resources we have access to, the optimum obtained for $minSim = 0.80$. This reduced the database further to 492 tweets.

Some examples of tweets translated into English and their cosine similarities are shown in Table 1. The original tweets in Portuguese can be found in *[censored to avoid revealing the authors' identity. The link will direct to one of the authors GitHub repository.]*.

Using the $minSim = 0.80$ criterion and averaging the cosine similarities connecting pairs of users, we generated a user dataset comprising 56 users and 144 connections to represent the backbone of fake tweets. Figure 2 shows the distribution of

Sentence 1	**Sentence 2**	**Similarity**
@noahshuter @jdoriajr I don't doubt that the vaccine is good. I doubt it's this super interest just for it.	@momentsbrasil the vaccine that has no proven efficacy.	0.70
@ncajunior @mitags vaccine theoretically immunizes, but only medicine treats the disease. it's good to separate things so as not to confuse. there are a lot of people who got vaccinated while being infected and it got worse, some even died. even to get vaccinated you need to strengthen your immune system with the right medicines.	@taschnernatalia you take the vaccine you have to wear a mask, distancing, alcohol, you run the risk of having serious side effects and even dying. you take the covid kit you are immunized and you don't need to take a vaccine. simple	0.75
pfizer/biontech vaccine approved for teenagers from 12 years old in the united states, which will allow the vaccination of the scoundrel before the next school year	From now on, teenagers aged 12 to 15 can get vaccinated against the coronavirus in the United States. the authorities in the country authorize the use of immunizer from pfizer/biontech after a study showed that the vaccine is 100% effective against the disease for people in this age group.	0.80
The phenomenon of the sinovac vaccine is strange, countries report an increase in cases after the use of the vaccine! today's most important thread!	the strange phenomenon of the sinovac vaccine: countries report increase in cases after using the vaccine #equipejd @equipejd	0.85
south africa suspends astrazeneca vaccine after evidence of limited protection against variant. according to the minister of health south african, data from a clinical trial show that the astrazeneca vaccine offers limited protection against the variant of coronavirus	south africa suspends vaccine from astrazeneca due to low efficacy against coronavirus variant. inumizing agent showed limited protection against the disease caused by the variant of the dominant coronavirus in the country.	0.90
china admits low effectiveness of coronac eunews- in a moment of transparency unusual, the director of the center of chinese disease control gao fu, confirmed, the day before, the low effectiveness of coronavac, the vaccine	china admits the low effectiveness of the coronavac. in a moment of transparency unusual, the director of the center of Chinese disease control gao fu, confirmed, a day before, the low effectiveness of the coronavac, the vaccine produced by the pharmaceutical company in the country	0.95

Table 1: Examples of tweets translated to English and their cosine similarity.

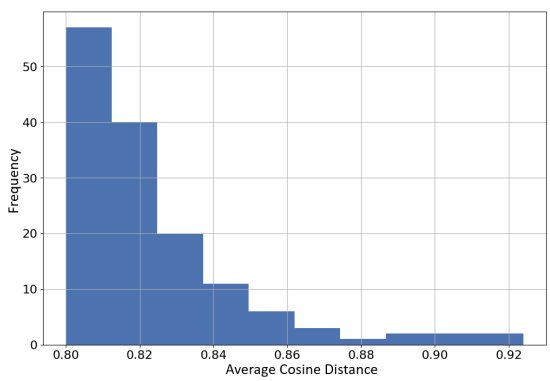

Figure 2: Weights connecting spreaders of fake tweets. The average of cosine similarities was used whenever the users were connected by more than one similar tweet.

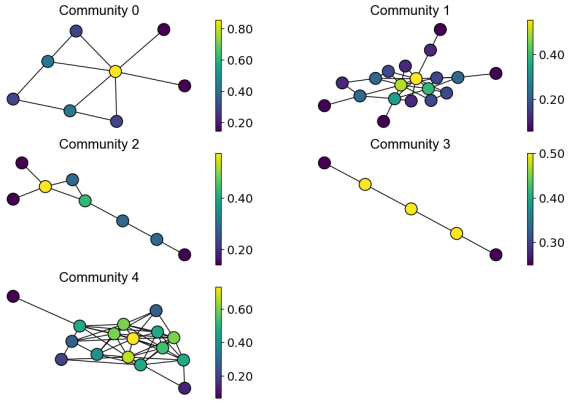

Figure 4: Degree Centrality of each community.

weights in the user dataset. The average cosine distance distribution decays rapidly up to approximately 0.87, then stays relatively flat with 1 and 2 interactions only. The highest value for the average cosine distance is 0.92.

The resulting graph is fully connected, i.e., every user is connected to at least one other user. It comprises a strongly clustered and a sparse regions with very few degree-1 nodes (users connected to only one other user). Applying the Louvain algorithm to optimize modularity revealed 5 communities with 0.44 modularity, as represented in Figure 3. The degree of centrality for each community is shown in Figure 4. Except for community 3, each community has a clear central user who is the primary source of fake tweets.

The processing of the dataset is summarized in the flowchart in Figure 5.

To test for consistency and understand the communities, the authors scrutinized all tweets related to each community. This also allowed us to determine their central topic. Examples of tweets in each community are in Table 2.

Table 3 shows the size and main subject of each community. Community 1 and 4 are the largest, with 19 and 16 members respectively. Communities 0 and 4 have a few false positives.

Figure 4 shows the degree of centrality of each community.

Community 0 was related to fake tweets concerning Instituto Butantan, the research center that developed the *Coronavac* vaccine, in cooperation with the Chinese company Sinovac. Instituto Butantan belongs to the State of São Paulo, whose governor at the time opposed the federal government. The fake tweets were mainly related to the reliability of the vaccine, insinuating poor quality due to its Chinese origin. Surprisingly, the central user of community 0 is the official account of Instituto Butantan, with 0.86 centrality. This was not a source of disinformation. However, the remainder of users in this community tweeted vaccine disin-

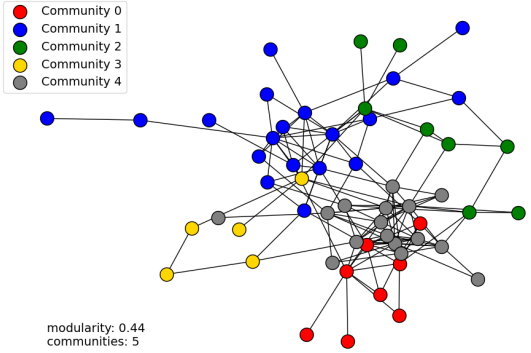

Figure 3: Community structure of the graph. There are five communities with a 0.44 modularity.

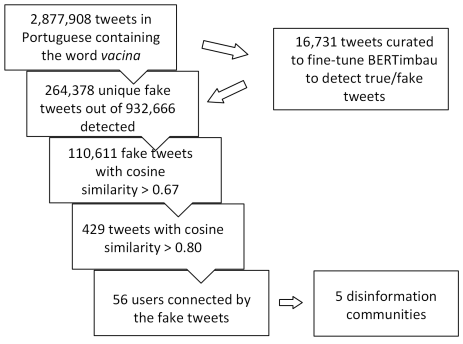

Figure 5: Summary of the data processing used in this work. The 5 disinformation communities are found from the 56 authors of 429 fake tweets with cosine similarity > 0.80.

Community	Text
0	Butantan only delivered the pre-clinical studies and not the clinical ones conducted in humans. Sinovac doesn't even have certification for good manufacturing practices. A safe vaccine typically takes about 7 years to be developed. There's no problem with that? Only for the servants of China.
1	But a Brazilian took this vaccine and died, I wouldn't take this vaccine myself since you guys are so crazy, go ahead
2	You said everything. Brazil already producing the vaccine and a second wave ravaging Europe. Either we are pioneers or this vaccine is shit that no one in Europe wants and it doesn't work. For me it's all speculation and deception. Wake up, a vaccine won't be produced in months.
3	Do you understand? Did @jdoriajr, aka the half-assed dictator, also understand or is he still in the mood to apply a vaccine that has no proof? Or just buy it and not use it!
4	the truth about coronavirus. understand the technology used by Sinovac, the risks involving the Chinese vaccine and the real scenario of vaccine production.

Table 2: Examples of tweets for each community. The text was translated from Portuguese trying to preserve grammatical errors and expressions.

Community	Members	Main subject
0	8	Instituto Butantan
1	19	Adverse effects of the vaccines
2	8	Cast doubt on vaccine efficacy and safety
3	5	Government and governors
4	16	News

Table 3: Communities breakdown. The main subject of each was determined by reading the associated tweets. We did not find any false positives in the classification, except in communities 0 and 4.

formation. It turned out that the name Butantan was cited in many fake tweets from anti-vaxers and vaccine deniers. The algorithm used to detect fake tweets ended up associating the word Butantan with fake tweets.

Community 1 was related to spreading fake tweets that denounced the supposed dangerous adverse effects of the vaccine, such as myocarditis, strokes, and sudden death. The tweets in this community had the intention of inducing vaccinal hesitation. The most central user has been banned from Twitter. The second most central user has also been banned from Twitter.

Community 2 involved fake tweets to cast doubt on vaccine efficacy and safety. They claimed the vaccine would not prevent contagion or protect against the disease. This is a more linear community with only two members having more than two connections. The most central user has been banned from Twitter.

Community 3, the smallest, was about fake tweets criticizing state governors who opposed the federal government, especially the one of São Paulo State. These tweets combined fake information about vaccines with political criticism. This community does not have a prominent central node, and its structure is linear, without interconnected nodes. One of its three most central users has been banned from Twitter.

Community 4 is about spreading fake tweets about vaccines and conspiracies involving news networks. The Brazilian anti-vax movement used to accuse the press of falsifying information to hide the supposed bad side effects of the vaccines. Its most central user has been banned from Twitter. The second most central user is BandNews, a reputed radio and TV channel, which is obviously a false positive. Again, this channel was so cited in fake tweets that the classification algorithm associated it with fake tweets.

Figure 6 shows the degree of centrality, the closeness centrality, and the eigenvector centrality for the whole graph.

The eigenvector centrality shows that the users around the nodes with the highest centrality in the graph that belong to Community 1 are also influential in the network. This is not the case for the remainder of the communities.

The node with the highest degree of centrality belongs to community 4. It has 15 edges out of a total of 144. Twitter has suspended this user. This user was still suspended in September 2024, indicating it was probably permanently banned. Although X does not publicly discuss the reasons for banning an account, it was a disinformation spreader. Before the acquisition of Twitter and its name change to X, the community guidelines stated that: "We may suspend an account if it has been reported to us as violating our Twitter Rules surrounding abuse. When an account engages in abusive behavior, like sending threats to others or impersonating other accounts, we may suspend it temporarily or, in some cases, permanently." (Twitter, Accessed: 11/15/2022).

Seven out of the ten most central users in the graph have been suspended or permanently banned from Twitter.

5 Conclusions

In this work, we used semantic proximity as a criterion to identify the formation of disinformation communities on Twitter (currently X). We studied the specific case of fake tweets about COVID-19 vaccines in Portuguese.

We used a curated subset of all tweets on the subject over a certain period to fine-tune BERTimbau, the Brazilian Portuguese version of BERT. Taking advantage of the transformer-based embedding of BERTimbau, we used semantic proximity between fake tweets, measured by their internal product, to estimate engagement between users and the formation of disinformation communities. Using graph theory, we could detect the formation of five communities, each with well-defined central users (except for the more general News community with more diffuse central users). Most central users have been banned from Twitter for violating community guidelines[4]. This work shows that semantic proximity between texts is a very useful criterion for detecting disinformation communities in social networks.

6 Limitations

We used a fine-tuned BERTimbau framework to obtain the fake tweets subset. Even taking measures to mitigate dataset imbalance, the algorithm is still biased and may classify real tweets as fake.

Due to computational limitations, only a reduced subset of fake tweets, corresponding to tweets with higher semantic proximity, were considered. Only communities with highly similar tweets, with

[4]The enforcement of community guidelines has been relaxed after Twitter acquisition and name change to X.

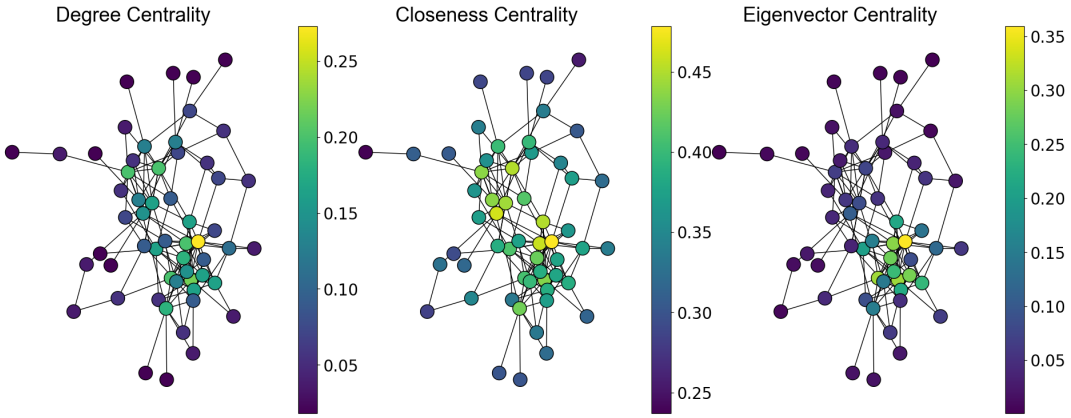

Figure 6: **Left:** Degree centrality. **Center:** Closeness centrality. **Right:** Eigenvector centrality.

$minSim$ greater than 0.80, were studied. Many less connected communities may exist, but they would not have been detected here.

Although the Louvain algorithm is very efficient in detecting communities, it forces a node to be a member of a single community. Users are likely to participate in more than one community.

References

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Wilson Ceron, Gabriela Gruszynski Sanseverino, Mathias-Felipe de Lima-Santos, and Marcos G. Quiles. 2021. COVID-19 fake news diffusion across Latin America. *Social Network Analysis and Mining*, 11(1):47.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10.

Roger C Entringer, Douglas E Jackson, and DA Snyder. 1976. Distance in graphs. *Czechoslovak Mathematical Journal*, 26(2):283–296.

Cláudia Pereira Galhardi, Neyson Pinheiro Freire, Maria Cecília de Souza Minayo, and Maria Clara Marques Fagundes. 2020. Fact or fake? An analysis of disinformation regarding the Covid-19 pandemic in Brazil. *Ciência & Saúde Coletiva*, 25:4201–4210.

Rafael Geurgas and Leandro R Tessler. 2024. Automatic detection of fake tweets about the covid-19 vaccine in portuguese. *Social Network Analysis and Mining*, 14(1):55.

Mark E J Newman. 2018. *Networks*. Oxford university press.

Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. 2024. Reuters Institute digital news report 2024. Technical report, Reuters Institute for the Study of Journalism.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2018. "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. *arXiv preprint*. ArXiv:1801.00317 [cs].

Julie Ricard and Juliano Medeiros. 2020. Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, 1(3).

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Twitter. Accessed: 11/15/2022. Help on your suspended twitter account.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.