Investigating the Use of Negative Speech by Bots in Brazilian Political Discourse

Kai Barber-Harris

November 2023

Abstract

In recent years, the digitisation of the world has given rise to an increasingly polarised sociopolitical climate, especially in online spaces such as those that exist on X, formerly Twitter. In this literature review we cover and summarise an extremely wide variety of previous work already existing in political and linguistic language processing, bot detection, and political alignment classification to support the planned project to investigate how bots may be negatively affecting these online spaces and encouraging the use of negative language in general or hate speech toward minority groups. We hope to extend the research into why these spaces have become more hostile, and plan for all tools used for this analysis to be free and open-source as part of our contribution to the discussion.

I certify that all material in this dissertation which is not my own work has been identified.

Signed:

1 Introduction

As the world has become increasingly digitised, gigantic corporations formed around the existence of a social media website have become commonplace. One of these sites is X, formerly Twitter, which exists as a microblogging service most commonly used to rapidly share information and disseminate thoughts through a populace. A common concern of these platforms is the existence of automated accounts, colloquially known as bots, and their influence on online conversations. While this was not always an area that was seen as important or even interesting, the number of media outlets and researchers covering bots online has exploded since the 2016 US presidential elections and the UK Brexit vote of the same year. More recently, the 2020 Covid-19 pandemic caused discord and distrust in major news outlets and government figures, leading people to social media for their news especially in the younger generation (Shearer 2023). As the pandemic developed, the discussion became partisan, and prone to manipulation by bots as a sociopolitical issue rather than a scientific one.

The average person using social media, through these events, has become more aware than ever of the potential effects of external influences on discourse and conversation online. A vast corpus of work has accumulated since these events first sparked major concern, and this literature review hopes to summarise the existing relevant content in an easily digestible way.

To that end, as it has been shown in these events that bots can have a serious and dangerous effect on the political landscape online, we noticed that no study seems to have considered how these bots might be affecting the tone and method of communication online. We believe that it may be possible to study the perceived increased polarisation of politics in recent years and correlate it to bots encouraging the use of negative language, derogatory statements, and possibly even hate speech targeted at specific groups of people.

The data to be used for the project is the same as used by Pacheco (2023), and comprises 437 million X posts by 13 million accounts spanning nearly 5 years of collection, with the vast majority of the corpus written in Portuguese and concerning politics in Brazil. See section 3.5 for more information.

A thorough analysis of this data with the goal of quantifying the effect bots have on the use of language and tone online is crucial to understanding the increasingly polarised landscape of discussion, both online and in real life. A correlation in any direction between bot activity and this polarisation would have wide-reaching effects on the further study and consideration of the world's political systems in future. Confirmation of bot activity influencing this phenomenon would add to the already gathered evidence of the necessity of online platforms to implement better bot-detection tools and stricter guidelines for eliminating them from the conversation, along with allowing a new lens of examination to researchers studying why these systems have become so difficult to navigate.

2 Existing Literature

2.1 Sentiment Analysis

This subsection concerns sentiment analysis, and the large body of previous work performed on its calculation and analysis since Pang, Lee, and Vaithyanathan (2002) became one of the earliest examples of computer-assisted sentiment analysis. More recently, Zimbra et al. (2018) performed a survey on the state-of-the-art in sentiment analysis for X data. Zimbra et al. identify that domain-specific models trained on relevant data outperform generalist models by 11% on average. More importantly, this paper found that the performance of sentiment analysis tools was "lacklustre", with an average accuracy of 61% overall and a large range in accuracy of 31%. The paper also finds that BPEF, NRC and Webis are the most effective classification techniques, boasting a 71% accuracy across the paper's test data covering 5 different domains. BPEF also had a slightly better recall rate than the other options.

An earlier survey by Giachanou and Crestani (2016) notes that sentiment analysis is “an open domain” and does not have many tools effective for multilingual content. This is especially relevant for this review, as the data planned for use is mostly in Portuguese. Notably, some work does exist in this subject. Especially relevant is Tumitan and Becker (2013)’s work on classifying sentiment in comments on a Brazilian news site, most of which is in Portuguese. Tumitan and Becker deploy a system leveraging SentiLex-PT, developed by Teixeira et al. (2008) in their book on studying computational processing of Portuguese. Much more recently, and equally important, Melo and Figueiredo (2021) processed news articles written in Portuguese for a Brazilian audience using the spaCy2 and VADER Python libraries to calculate text sentiment in a language-agnostic manner which would be just as effective on Portuguese as English. Finally, Barhan and Shakhomirov (n.d.) contains detailed explanations of each individual step of performing a sentiment analysis, and presents a system for automatic analysis of sentiment in X messages.

These papers and results together will allow us to develop our own method for sentiment analysis built on the techniques already found to work, most likely using the VADER and spaCy2 Python libraries.

2.2 Stance Analysis

Stance analysis as a field has existed for some time; Bestvater and Monroe (2023) provides a short history, noting that stance has been discussed linguistically since Biber and Finegan (1988) defined it as “the expression of a speaker’s standpoint and judgment toward a given proposition”. However, calculating stance in an automated manner is much more of a conceptual problem as compared to sentiment calculation.

Due to the need to determine the meaning of words in context to successfully determine stance, deep learning approaches suit the task well along with feature-based machine learning techniques such as naive Bayes, logistic regression, and support vector machines (Küçük and Can 2020). Pretrained transformer models are also becoming more common, with Bestvater and Monroe (2023)’s analysis depending mostly upon BERT (Bidirectional Encoder Representations for Transformers) for their stance detection, a model introduced by Vaswani et al. (2023) and implemented with the Simple Transformers Python library (Rajapakse (2023), Wolf et al. (2020)). Another model recently presented comes from Kucher et al. (2020), which developed a tool called StanceVis Prime. Kucher’s model is explained in the paper, and is formed of the VADER sentiment classifier and “a custom logistic-regression based stance classifier”, the latter of which is implemented with the scikit-learn Python library.

ALDayel and Magdy (2021) performed a survey of stance detection methods and their workings. This survey splits stance detection methods into 3 broad categories: supervised machine learning, unconstrained-supervised machine learning, and unsupervised machine learning. The survey notes that supervised machine learning is by far the most common choice in the field; a list of six citations follows this claim. In supervised machine learning stance analysis, the goal is to annotate a stance dataset with a predefined set of labels, most commonly either two antonymous terms or three, formed of two antonymous terms with a neutral option (ALDayel and Magdy 2021).

The work goes on to explain unconstrained supervised learning. The motivation to use unconstrained methods such as transfer learning and weak supervision is the “scarcity of labelled data”. Unconstrained supervised methods share a trait in that they can obtain knowledge from one data set and apply it to another, alleviating the problem of requiring large amounts of labelled data to train from. While this can appear to be purely a benefit, it stands to reason that Zimbra et al. (2018)’s findings about model accuracy would still hold, as stance analysis is a subdomain of sentiment analysis. That is, a model trained on even a small amount of relevant data is almost always more accurate than a generalist model trained on other data. Finally, unsupervised learning for stance detection is a field that has begun receiving attention as a serious methodology. Unsupervised methods generally group input data in such a way that manual review can assign categories to the identified clusters once processing has completed.

ALDayel and Magdy note that "clustering techniques are primarily used with a focus on user and topic representation" for unsupervised stance detection. All facets of collected data can be used in this form of classification to boost model performance; Darwish et al. (2020) found that "using retweets as a feature provided the best performance score upon implementing clustering algorithm (DBSCAN), which surpassed the supervised method when using the fast-text and SVM models." (ALDayel and Magdy 2021). Further, ALDayel and Magdy's survey covers Rashed et al. (2021), which "introduced embeddings representations of users' tweets to enhance [their] stance detection model." This improvement allowed Rashed et al. to analyse "fine-grained polarisation" between politically-themed X posts. Finally, Qi (2023) extends the work completed by Bestvater and Monroe (2023), proposing the inclusion of a topic metric to accompany stance and sentiment analysis. This methodology was found by Qi to increase political stance classification by 19% when compared to just using sentiment alone.

These findings, considered in parallel with the earlier work on sentiment analysis, will aid us in our future work to create our own effective stance analysis model for a deeper look into the data beyond sentiment alone.

2.3 Hate Speech Analysis

Analysis of hate speech is a popular area of study for linguistic computing researchers as social movements around the globe have become more concerned with whether this behaviour online may be able to predict events such as hate crimes or terrorist activities. The current canonical survey paper in this area is Ayo et al. (2020), which details a variety of machine-learning techniques for the classification of hate speech from X data. Ayo's work differentiates between two categories of model: "single" methods which apply a single machine learning classifier, and "hybrid" methods which apply multiple, and are "more computationally efficient and produce superior results than their single method counterparts.".

An important extension from Ayo et al's survey is Pereira-Kohatsu et al. (2019), which analyses a wide variety of alternate methods of classification, and even presents a language-agnostic system that outperforms all of the surveyed methods for large scale hate speech analysis. Pereira-Kohatsu also finds that embedding-based models tend to perform better than frequency-based models, and for these embedding models including point-of-speech tags and suffixes do not significantly impact classification accuracy.

Finally, Watanabe, Bouazizi, and Ohtsuki (2018) provides a method of hate speech classification that reaches an accuracy of 87.4% for binary classification, and a slightly lower accuracy of 78.4% for ternary classification. Watanabe's approach is based on the use of unigrams, sentiment, and semantic language patterns to determine whether the X post is hate speech or not. The paper describes in detail how each of these features in a text can contribute to hate speech detection. None of these methods or techniques are unique to the English language and so can likely be adapted to our use case in Portuguese.

These papers together provide a comprehensive overview of how we might approach identifying hate speech in our own study, including the complication of the data being written in Portuguese.

2.4 Political Alignment Classification

A variety of work already exists on the topic of identifying the political alignment of a user on social media, as it is a common wish for researchers to study the notable divide which has formed in political discussions over time.

Most immediately relevant is Ansari et al. (2020), which examines political discourse in India on X. Ansari et al. explain the use of common machine learning techniques to identify the political leaning of a user, finding that a Long Short Term Memory (LSTM) model under a trigram implementation worked the best in their data.

Though Ansari et al's work is recent, it leaves a want for detail. Barberá et al. (2015) provides this detail, explaining in depth a process they developed to estimate ideological placement of X users in the United States. The method involves projecting the "ideological coordinates" of users into a plane based on the political accounts they follow along with the political accounts themselves. By examining the other accounts nearby any given user, we can approximate their ideological beliefs. This algorithm could easily be extended with k-means clustering to return a discrete classifier between political alignments.

These two papers are detailed and thorough; the work contained within them is more than enough to kickstart the development of our own classifier for use in our analysis.

2.5 Bots and Coordinated Action

As social media has become more and more important in our day-to-day lives, the importance of being able to identify coordinated actors and non-human interactions has only increased. Because of this, much work has been completed in the detection of both of these phenomena, especially in data collected from X.

Pacheco (2023) is his most recent and most in-depth study, working on the same dataset he has provided us access to. Pacheco's analysis revealed that since collection began, the rate of bot engagement has grown "alarmingly", with various metrics contributing to this conclusion.

A variety of recent papers are more specifically concerned with the spread of disinformation and how bots may contribute to this phenomenon. Vosoughi, Roy, and Aral (2018) is one of the leading papers in this concept, showing definitively that "fake news" does indeed traverse the information space on X much faster than factually correct news, but also that bots are not responsible for this effect on a large scale; in fact, Vosoughi et al. finds that bots tend to spread false and true news at the same rate overall. On the other hand, Shao et al. (2018) performs a slightly more granular study on the same topic, showing that bots are instrumental to the amplification and therefore the spread of messages from low-credibility sources. The paper details how bots are responsible for early bursts in engagement on low-credibility messages, adding to a post's ability to go viral online and reach a much larger audience.

It then stands to reason that considering both Vosoughi et al. and Shao et al.'s findings, despite bots overall boosting low and high-credibility content roughly evenly, low-credibility content utilises bots in a more targeted, effective manner that further potentiates the ability of bots to affect conversations online. This theory is in fact supported by the findings of Stella, Ferrara, and De Domenico (2018), which shows definitively that bots exist and act in political ecosystems online in a focused manner, interacting with and targeting specific groups of users that fall into categories presumably set by the bots' controllers.

Finally, Starbird, Arif, and Wilson (2019) demonstrates the fact that bot activity online is both influenced by and influences organic human activity; the connection intuitively may be assumed to be that bots influence humans, but in fact humans also appear to influence bot behaviour. Though the exact reason for this is unclear, it has interesting implications in the study of echo chambers and bot activity online; we cannot simply consider the actions of bots alone, and must also acknowledge that the actions of the humans surrounding bots may change the influencing behaviour of the bots themselves.

These works will greatly assist us in developing our understanding of our data as we begin to analyse it and correlate the previous sections of our work to a user's botscore, which is already included in our data (see sections 3.3 and 3.5 for more information).

2.6 Framing

The framing of a text is a tangentially related field to the subjects already discussed in earlier sections. While not essential, this section explains the concept of framing in an analysis context and why it is important to acknowledge its potential use in our work.

Kwak et al. (2021) recently published an extensive and extremely poignant paper explaining the importance of framing in language processing, and presented an unsupervised system named FrameAxis to tease out microframes in large data sets. A microframe is an axis of phrasing, such as "legal/illegal" or "appealing/unappealing". The authors also note the ability for FrameAxis to be fed predetermined microframes to work from "when authors are already aware of important candidate frames of the text", which can be especially useful for political applications such as ours.

The motivation for including such an analysis, as phrased by the authors, is that "by focusing on a particular aspect over another, even without making any biased argument, a biased understanding of the listeners can be induced." That is, a neutrally-sentimented and neutrally-stanced statement can still influence the perception of the reader on any given issue and therefore implicitly forward an unspoken agenda. FrameAxis can help detect this on a large scale in wide datasets in a language-agnostic manner.

The programming work performed by Kwak et al. and final product of FrameAxis is generously open source and available for use and review in a public github repository (Haewoon 2023). If the project is successful and the timeline allows for a framing analysis to be included, we will endeavour to ensure that this work goes ahead. More detail on this can be found in section 3.4.

3 Project Plan

Noting the existence of the literature just covered, we noticed what appears to be a gap in research. While studies exist covering a broad range of topics in the study and analysis of social media, including: sentiment analysis; stance analysis; hate speech classification; classification of users into political groups; framing and microframing analysis; the identification of users as bots; and the identification of coordinated actions online, no study or research appears to group these factors into one study to answer specific questions about the influence of bots and coordinated actions on political discourse and its general attitudes. Given these assets, along with the enormous data on Brazilian politics-focused X from August 2018 to March 2023, we feel well positioned to attempt to answer the following questions:

1. Are bots performing in a coordinated manner attempting to influence political discussions to be more incendiary?
2. Do coordinated activities or bot accounts admit a notable difference in average sentiment or stance toward a specific topic when compared to the generic real user base?

The remainder of this project plan outlines the planned methodology for performing analysis to answer these questions. The plan includes explanations of what work will be performed in the 6 months between now (November 2023) and the dissertation deadline (May 2024), including a Gantt chart summarising and visualising the planned timeline for the work.

We would consider this project to be successful if either one of the two questions presented has been definitively answered, or both questions have been explored and further areas of research have been identified to expand upon the findings. A definitive answer to both questions would be considered to be an outstanding success. We also hope that the open-source Python code developed in the course of this work will include a fully-passing testing suite on sample data and complete documentation of methodology and explanations of each internal process. These planned contributions together and the criteria set out in the project plan should allow a comprehensive postmortem to be completed in May 2024, including a return to this literature review and project plan to evaluate whether the questions put forward have been sufficiently answered and whether this was accomplished by utilisation of the project plan detailed in this section.

3.1 Ethical Concerns

It is fair and expected to raise the point of ethics concerns when speaking about a dataset such as our own. There are no direct ethics concerns with the data itself; the reason for collection was legitimate and was performed through the proper channels. However, we must be careful in the way we handle the data to avoid potential ethical issues.

Primarily, we note that no specific individual should be singled out during this investigation. There is no reason to examine individual users in this particular analysis; we are focusing on the high-level potential effects of a bot population acting within human conversations. Naturally following from this, there is no reason to write about any single user in the final project. In some situations, it may be reasonable to examine the network surrounding an organisation or public figure in the same manner that Pacheco, Flammini, and Menczer (2020) did, but this is not the same as examining a single personal account.

Beyond this, we cannot identify any other notable ethical concerns for this project. The data is pseudoanonymous by nature, and no analysis will be performed which could appear dangerous or otherwise revealing to the related users.

3.2 Deliverables

The primary deliverable of this project will be the final dissertation to be handed in May 2024. This dissertation will include the final written analysis of the work completed, including relevant graphs and figures drawn to improve reader understanding, and will aim to answer the questions put forward in section 3. Also deliverable will be the code used to perform this analysis and all its constituent modules (section 3.3). We will host this Python code as open-source on github.com, a popular site used for hosting and sharing code. It is our hope that this code will be of a sufficient standard and documentation that future research in the field may be able to leverage it for their own projects, qualifying it as a worthwhile contribution to the wider discussion alongside the final write-up.

3.3 Project Modules

Writing, testing, and documenting the code to perform this analysis is expected to be the primary use of time during this project. Due to the varied analysis planned, we expect to write a variety of Python scripts which each perform their own function in the analysis, and then write a connector script which puts them all together once each has been completed and validated. In order, these are the planned modules:

1. Sentiment analysis script.
2. Stance analysis script.
3. Hate speech classifier script.
4. Political alignment classifier script.
5. Bot detection script.
6. FrameAxis adaptation. (Project extension. See section 3.4)
7. Module connector.

Implementation of each of these scripts is supported by the literature covered in the first section of this writing (section 2). As such, the development process of the algorithms should be simplified; the conceptualisation and evaluation of the methods has already been done for us, and all that remains is to implement each for our purposes. Work that has already been completed toward these goals includes:

1. Preprocessing of the data for analysis, performed by Pacheco.
2. Precalculation of botscores in data as part of this preprocessing.
3. Creation of a frame analysis algorithm (FrameAxis).

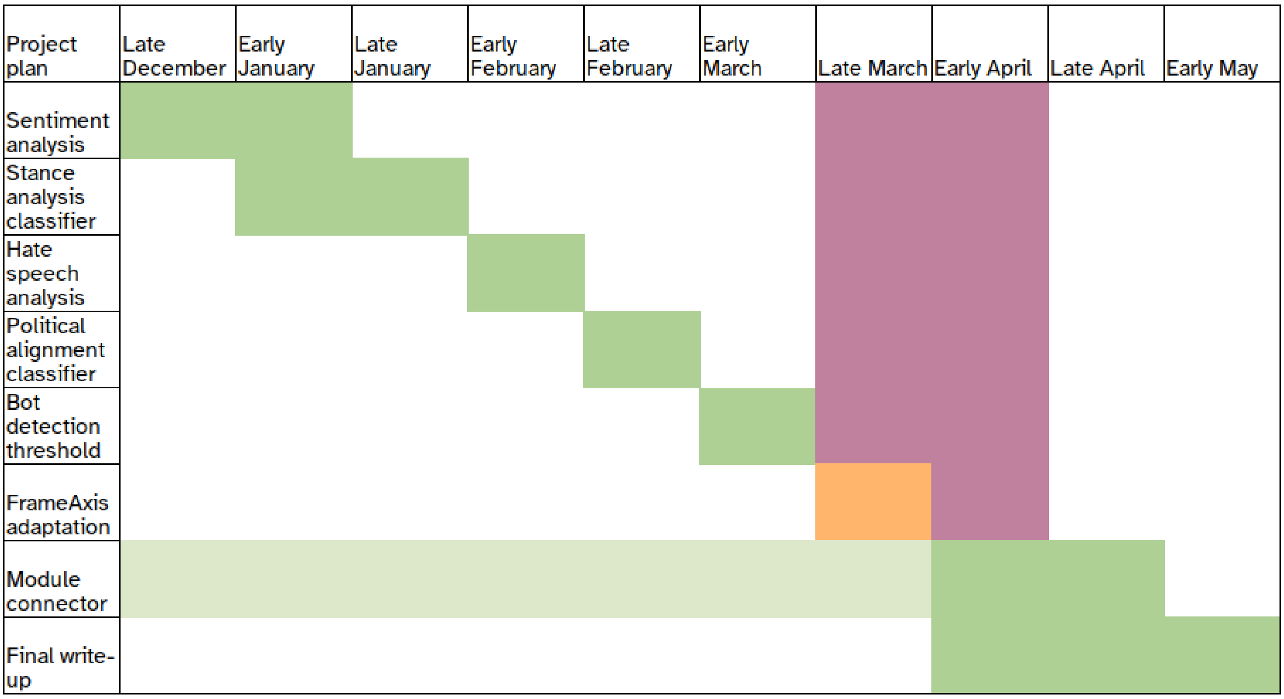
Our data already being preprocessed saves much time for us, but the data will still need to be validated as suitable for these purposes. The precalculation of botscores saves us from having to do it ourselves; this botscore is a measure of how likely a user is to be a bot based on their activity. We must still calculate an appropriate botscore threshold to consider a user as a bot in our work. Likewise, FrameAxis is generously provided as open-source by Kwak et al. (2021), minimising the work needed to be completed for this step. It is likely that a connecting program will need to be written regardless.

The module connector to be written will be a simple Python script which calls each of the previous modules in sequence and stores the data in such a way that it is easy to comb through and work with for analysis and the writing of the final project. This connector will also be the point at which we generate any graphs or figures required for the final write-up. We will be working on this script periodically as the other modules are created; the final block of time dedicated to the connector will be time specifically used to make sure the data is easy to tabulate and analyse.

3.4 Project Management

This section includes a Gantt chart to detail the project timeline. Each column represents a two week block, generally the first and second half of each month. In early January, sentiment and stance analysis overlap as stance analysis leverages sentiment analysis methods, and the two should be considered together.

We also actively acknowledge that there is a lot to do in this project. We do not wish to commit fully to an overextension and end up falling short. To this end, we have included the two purple columns in late March and early April as a mitigation strategy. These columns represent time to return to earlier sections if needed; otherwise, it will be dedicated to the implementation of a framing analysis, represented by the orange block in late March, and a good headstart on the module connector and final write-up in early April. This can be considered an extension to the project if all other work is



Project plan	Late December	Early January	Late January	Early February	Late February	Early March	Late March	Early April	Late April	Early May
Sentiment analysis	🟩	🟩					🟪	🟪		
Stance analysis classifier		🟩	🟩				🟪	🟪		
Hate speech analysis				🟩			🟪	🟪		
Political alignment classifier					🟩		🟪	🟪		
Bot detection threshold						🟩	🟪	🟪		
FrameAxis adaptation							🟧	🟪		
Module connector	▨	▨	▨	▨	▨	▨	▨	🟩	🟩	
Final write-up								🟩	🟩	🟩

Figure 1: The Gantt chart detailing the project plan and the timeline for each section of the project.

completed to a sufficient standard. The work involved is non-essential to our investigation, but would serve to improve and deepen the study if it could be included.

3.5 Data

The data we will be working with is provided by Diogo Pacheco, who has an extensive backlist in this field (Pacheco (2023), Pacheco, Flammini, and Menczer (2020), Chen et al. (2021)). Pacheco's dataset consists of 437 million individual X posts made by 13 million unique accounts, collected via the X streaming API from August 2018 until the API was closed by the new administration of X in March of 2023.

The data was collected through a set of keywords themed around Brazilian politics; the set also includes X posts from candidate accounts, X posts containing the official hashtag of each candidate's campaign, and X posts containing the full name of a candidate. Across the collection period, the keywords were only adjusted once, in July 2022 (Pacheco 2023). This adjustment also marks the start of the collection of X posts from official party accounts until the end of data collection. Due to the subject matter of the data collection, the vast majority of the dataset is in Portuguese, the primary and national language of Brazil. This will present some challenges later working with the data, as we do not speak Portuguese. Pacheco provides an already preprocessed version of the data, including embedded and pre-calculated "botscores" generated by the now-defunct Botometer tool, a casualty of the changes to X's APIs earlier this year.

Brazil is the 4th-largest user of X in the world (*Countries with Most X/Twitter Users 2023* 2023), with more than 24 million active accounts. Further, over the last 5 years, Brazil has undergone many major political events; examples which have been tracked in our data include an attempted coup, a presidential and a local election cycle, and the general day-in day-out discourse of modern online politics. These factors together lend our data to a varied and detailed analysis of political discourse both in times of stability and in times of uncertainty and unrest.

References

ALDayel, Abeer and Walid Magdy (July 2021). “Stance Detection on Social Media: State of the Art and Trends”. In: *Information Processing & Management* 58.4, p. 102597. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102597. (Visited on 11/13/2023).

Ansari, Mohd Zeeshan et al. (Jan. 2020). “Analysis of Political Sentiment Orientations on Twitter”. In: *Procedia Computer Science.* International Conference on Computational Intelligence and Data Science 167, pp. 1821–1828. ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.03.201. (Visited on 11/07/2023).

Ayo, Femi Emmanuel et al. (Nov. 2020). “Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-of-the-art, Future Challenges and Research Directions”. In: *Computer Science Review* 38, p. 100311. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2020.100311. (Visited on 11/07/2023).

Barberá, Pablo et al. (Oct. 2015). “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?” In: *Psychological Science* 26.10, pp. 1531–1542. ISSN: 0956-7976. DOI: 10.1177/0956797615594620. (Visited on 11/07/2023).

Barhan, Anton and Andrey Shakhomirov (n.d.). “Methods for Sentiment Analysis of Twitter Messages”. In: ().

Bestvater, Samuel E. and Burt L. Monroe (Apr. 2023). “Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis”. In: *Political Analysis* 31.2, pp. 235–256. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2022.10. (Visited on 11/13/2023).

Biber, Douglas and Edward Finegan (Jan. 1988). “Adverbial Stance Types in English”. In: *Discourse Processes* 11.1, pp. 1–34. ISSN: 0163-853X. DOI: 10.1080/01638538809544689. (Visited on 11/14/2023).

Chen, Wen et al. (Sept. 2021). “Neutral Bots Probe Political Bias on Social Media”. In: *Nature Communications* 12.1, p. 5580. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25738-6. (Visited on 11/06/2023).

Countries with Most X/Twitter Users 2023 (2023). https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. (Visited on 11/18/2023).

Darwish, Kareem et al. (May 2020). “Unsupervised User Stance Detection on Twitter”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 14, pp. 141–152. ISSN: 2334-0770. DOI: 10.1609/icwsm.v14i1.7286. (Visited on 11/16/2023).

Giachanou, Anastasia and Fabio Crestani (June 2016). “Like It or Not: A Survey of Twitter Sentiment Analysis Methods”. In: *ACM Computing Surveys* 49.2, 28:1–28:41. ISSN: 0360-0300. DOI: 10.1145/2938640. (Visited on 11/07/2023).

Haewoon, Kwak (Apr. 2023). *Frameaxis.* (Visited on 11/14/2023).

Kucher, Kostiantyn et al. (Dec. 2020). “StanceVis Prime: Visual Analysis of Sentiment and Stance in Social Media Texts”. In: *Journal of Visualization* 23.6, pp. 1015–1034. ISSN: 1343-8875, 1875-8975. DOI: 10.1007/s12650-020-00684-5. (Visited on 11/13/2023).

Küçük, Dilek and Fazli Can (Feb. 2020). “Stance Detection: A Survey”. In: *ACM Computing Surveys* 53.1, 12:1–12:37. ISSN: 0360-0300. DOI: 10.1145/3369026. (Visited on 11/15/2023).

Kwak, Haewoon et al. (July 2021). “FrameAxis: Characterizing Microframe Bias and Intensity with Word Embedding”. In: *PeerJ Computer Science* 7, e644. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.644. (Visited on 11/13/2023).

Melo, Tiago de and Carlos M. S. Figueiredo (Feb. 2021). “Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach”. In: *JMIR Public Health and Surveillance* 7.2, e24585. DOI: 10.2196/24585. (Visited on 11/07/2023).

Pacheco, Diogo (Oct. 2023). *Bots, Elections, and Controversies: Twitter Insights from Brazil's Polarised Elections.* DOI: 10.48550/arXiv.2310.09051. arXiv: 2310.09051 [cs]. (Visited on 11/06/2023).

Pacheco, Diogo, Alessandro Flammini, and Filippo Menczer (Apr. 2020). "Unveiling Coordinated Groups Behind White Helmets Disinformation". In: *Companion Proceedings of the Web Conference 2020*. Taipei Taiwan: ACM, pp. 611–616. ISBN: 978-1-4503-7024-0. DOI: 10.1145/3366424.3385775. (Visited on 11/06/2023).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (May 2002). *Thumbs up? Sentiment Classification Using Machine Learning Techniques*. DOI: 10.48550/arXiv.cs/0205070. arXiv: cs/0205070. (Visited on 11/14/2023).

Pereira-Kohatsu, Juan Carlos et al. (Jan. 2019). "Detecting and Monitoring Hate Speech in Twitter". In: *Sensors* 19.21, p. 4654. ISSN: 1424-8220. DOI: 10.3390/s19214654. (Visited on 11/07/2023).

Qi, Weihong (Oct. 2023). *Beyond Sentiment: Leveraging Topic Metrics for Political Stance Classification*. arXiv: 2310.15429 [cs]. (Visited on 11/13/2023).

Rajapakse, Thilina (2023). *Simple Transformers*. https://simpletransformers.ai/. (Visited on 11/15/2023).

Rashed, Ammar et al. (May 2021). "Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey". In: *Proceedings of the International AAAI Conference on Web and Social Media* 15, pp. 537–548. ISSN: 2334-0770. DOI: 10.1609/icwsm.v15i1.18082. (Visited on 11/16/2023).

Shao, Chengcheng et al. (Nov. 2018). "The Spread of Low-Credibility Content by Social Bots". In: *Nature Communications* 9.1, p. 4787. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06930-7. (Visited on 11/07/2023).

Shearer, Elisa (2023). *Social Media Outpaces Print Newspapers in the U.S. as a News Source*. (Visited on 11/07/2023).

Starbird, Kate, Ahmer Arif, and Tom Wilson (Nov. 2019). "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, 127:1–127:26. DOI: 10.1145/3359229. (Visited on 11/07/2023).

Stella, Massimo, Emilio Ferrara, and Manlio De Domenico (Dec. 2018). "Bots Increase Exposure to Negative and Inflammatory Content in Online Social Systems". In: *Proceedings of the National Academy of Sciences* 115.49, pp. 12435–12440. DOI: 10.1073/pnas.1803470115. (Visited on 11/07/2023).

Teixeira, António et al., eds. (2008). *Computational Processing of the Portuguese Language: 8th International Conference, PROPOR 2008 Aveiro, Portugal, September 8-10, 2008 Proceedings*. Vol. 5190. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN: 978-3-540-85979-6 978-3-540-85980-2. DOI: 10.1007/978-3-540-85980-2. (Visited on 11/14/2023).

Tumitan, Diego and Karin Becker (2013). "Tracking Sentiment Evolution on User-Generated Content: A Case Study on the Brazilian Political Scene". In.

Vaswani, Ashish et al. (Aug. 2023). *Attention Is All You Need*. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. (Visited on 11/15/2023).

Vosoughi, Soroush, Deb Roy, and Sinan Aral (Mar. 2018). "The Spread of True and False News Online". In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559. (Visited on 11/07/2023).

Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki (2018). "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection". In: *IEEE Access* 6, pp. 13825–13835. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2806394. (Visited on 11/07/2023).

Wolf, Thomas et al. (July 2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. DOI: 10.48550/arXiv.1910.03771. arXiv: 1910.03771 [cs]. (Visited on 11/15/2023).

Zimbra, David et al. (Aug. 2018). "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation". In: *ACM Transactions on Management Information Systems* 9.2, 5:1–5:29. ISSN: 2158-656X. DOI: 10.1145/3185045. (Visited on 11/07/2023).