Natural Language Processing and Social Media: a systematic mapping on Brazilian leading events

Gabriele de S. Araújo[1], **Jéssica Brenda P. Leite**[1], **Marcelino S. da Silva**[1], **Antonio F. L. Jacob Junior**[2], **Fábio M. F. Lobato**[1]

[1]Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará – Santarém – PA – Brazil

[2]Departamento de Engenharia da Computação
Universidade Estadual do Maranhão – São Luís – MA – Brazil

antoniojunior@professor.uema.br, fabio.lobato@ufopa.edu.br

Abstract. *The number of social media platforms has increased significantly, as well as the number of active users. More than 18.2 million text messages are transmitted every minute on these platforms. Given the amount of data available, Natural Language Processing (NLP) techniques have been used by several researchers to analyze this large amount of unstructured data. Thus, it is essential to understand social media analysis's main trends and challenges. From this perspective, this study presents a systematic mapping of NLP for social media analysis considering papers published in five well-established academic Brazilian events: BRACIS, BraSNAM, ENIAC, STIL, and PROPOR. The study aims to identify the main tools and techniques used, tasks performed, data sources, and evaluation measures. For this purpose, 186 studies were analyzed and carefully selected among the 654 papers published in these events in the three years (2020 to 2022). The results show a glimpse of the current scenario on the subject and point out areas that can be improved in future research with techniques for tasks such as text classification, sentiment analysis, and named-entity recognition. Therefore, this work can be helpful for academics interested in exploring the potential NLP for social media analysis and having a clear view of gaps, challenges, and research opportunities in this area. Nevertheless, it should guide the productive sector in this knowledge transfer, reducing the gap between the state of the art and practice, consequently increasing the competitiveness and innovation of social media analysis tools.*

1. Introduction

Social media facilitates the connection between individuals and helps break down communication barriers, allowing everyone to share their stories and opinions [Hou et al. 2020]. [Kaplan and Haenlein 2010] describe social media as "*a group of applications based on the Internet and the ideological and technological foundations of Web 2.0 that allow the creation and exchange of User-Generated Content* (UGC)". In this sense, we can think of social media as the leading platforms and their features, such as Facebook, Instagram, and Twitter. In practical terms, we can also understand social media as an additional digital marketing channel that professionals can use to establish customer communication. From this perspective, social media becomes less about specific technologies or platforms and more about sharing information between users who have similar interests [Almeida et al. 2020, Appel et al. 2020].

Over the years, the number of social media sites and active users on these platforms has increased significantly, becoming one of the most important online applications [Aichner et al. 2021]. This fact has consequently led to the rise of communication via text, with over 18.2 million text messages transmitted every minute [Balaji et al. 2021]. The data generated by users have sparked academic interest, resulting in the increasing importance of the social media analysis field, which involves collecting and analyzing various social media data and extracting valuable and hidden information [Choi et al. 2020]. In the same direction, Natural Language Processing (NLP) has emerged as a promising approach to social media analysis. NLP is a subfield of Artificial Intelligence (AI) that uses computational techniques such as Machine Learning (ML) and Deep Learning (DL) for computers to learn, understand and produce human language content from the enormous amount of linguistic data available [Hirschberg and Manning 2015]. In this way, the intersection of these areas, combined with the ability of NLP to interpret and analyze linguistic data, drives the development of innovative approaches in social media analysis [Zhang and Lu 2021].

However, many challenges may depend on the natural language data context, making it difficult to achieve all goals with a single approach [de Oliveira et al. 2021]. For this reason, several researchers have widely studied the development of different tools and methods in the field of NLP, including specific tools and methods adapted to UGC [Khurana et al. 2023, Júnior et al. 2020]. With the growth of Brazilian communities of AI, data science, social media analysis [Lobato et al. 2021], and NLP, we wonder how knowledge in these areas is being spread across the scientific communities. To the best of our knowledge, there is no survey in the literature on methods and techniques for analyzing social media used in Brazilian events.

In order to fill this gap, we conducted a systematic mapping aiming to provide an overview of NLP techniques' application to social media analysis, identify the most used algorithms, and understand current trends in NLP use in this context. We have chosen the top five scientific events that publish work at the intersection of NLP and Social Media, namely: Brazilian Conference on Intelligent Systems (BRACIS), Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), *Encontro Nacional de Inteligência Artificial e Computacional* (ENIAC), Symposium in Information and Human Language Technology (STIL) and International Conference on Computational Processing of Portuguese Language (PROPOR). Our analysis included 654 papers from 2020 to 2022, of which 186 (30%) were reviewed. This period was chosen given the dynamic nature of the area, hence its rapid obsolescence.

The results obtained are helpful for researchers and practitioners interested in exploring the potential of these tools and techniques, having a clear picture of gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media. For the productive sector, this study can contribute to knowledge transfer, reduce the gap between the state of the art and practice, and increase the competitiveness and innovation of social media analysis tools. It is essential to point out that we are following Open Science Principles by providing all our data in a publicly available GitHub repository, allowing the reproducibility of the results.

The remainder of this paper is organized as follows: In Section 2, we discuss some related works. In Section 3 we present the systematic mapping protocols. The

results obtained are discussed in Section 4. Finally, the conclusion is given in Section 5, including directions for future research.

2. Related Works

Social media has become an essential data source for analysis across different sectors, including business, government, and the leisure industry [Hassani and Mosconi 2022, Mirzaalian and Halpenny 2019]. As the amount of data generated daily grows, data analysis techniques have become more important than ever in providing valuable insights [He et al. 2019]. Consequently, many researchers have explored this area to identify the nature of the data and the research domains addressed, thus allowing a better understanding of social events. Given what has been said, [Zachlod et al. 2022] conducted a systematic review of 94 papers that used or discussed social media data analysis as the main research topic between 2017 and 2020. This study identified that most of this data was collected from Twitter, Facebook, TripAdvisor, and LinkedIn.

Still in this perspective, [Khurana et al. 2023] aimed to present in detail the state of the art on trends and challenges in the field of NLP, with relevant works in the literature until 2022. It is seen that over the years, ML and DL methods have been used in different NLP tasks. Techniques used in multitasking learning are identified, such as Part-of-speech tagging (POS-tagging) and Named-Entity Recognition (NER), in Word Embedding, such as Global Vectors (GloVe) and Attention Mechanisms such as Transformers, with Bidirectional Encoder Representations from Transformers (BERT) being the most used. Despite significant advances, there are challenges; for instance, due to informal language, idioms, and culturally specific terms, there are few comprehensive linguistic models for different domains and geographic areas [Khurana et al. 2023, Pedroso et al. 2022].

[Souza et al. 2018] conducted a systematic mapping of studies related to the application of text mining to the Portuguese language from 1996 to 2014. The study used an automated search approach in digital libraries and a manual search in several conference proceedings held in Brazil (*e.g.*, PROPOR, BraSNAM, and STIL). Among the 234 tasks identified, text classification was the most addressed, representing 49% of the studies. The most used preprocessing technique for this task was stopwords removal, and the main algorithms employed were Support Vector Machine (SVM) and Naïve Bayes. Regarding data sources, approximately 50% of the studies were based on online news (*e.g.*, *Folha de São Paulo* and *Público*), while Twitter was identified as the main source of social media. Also, evaluation metrics, such as Precision, Recall, and F-measure, were prevalent.

In [Júnior et al. 2020], the author carried out a mapping of works published in international conferences of great impact in the area of data analysis from social media. The study adopted a systematic mapping approach, with targeted research questions to identify the most prevalent databases, tools, and algorithms in the studies, resulting in the analysis of 440 papers published between 2016 and 2019. The most widely used database was Twitter, followed by Facebook, Reddit, and Wikipedia. Regarding the tools, Linguistic Inquiry and Word Count (LIWC) were the most used, including Scikit-learn and Word2vec. As for algorithms, SVM, Logistic Regression (LR), and Long Short-term Memory (LSTM) were identified as the most frequent.

NLP and social media analysis still require a great deal of research and development, especially with the emergence of new platforms and the evolution of data analy-

sis models, as discussed in [Khurana et al. 2023]. Textual data analysis faces significant challenges, especially in the use of techniques and tasks specific to particular languages. The limitations of algorithms and tools for these languages are an important obstacle in this scenario. Likewise, few studies have focused on Brazilian events, as in the case of [Souza et al. 2018], whose mapping covered only up to 2014, and [Júnior et al. 2020] which was based on studies of international conferences. Therefore, the need for a systematic mapping directed to the NLP in social media analysis comes from the lack of works that show the state of the art focused on Brazilian academic events, in order to fill this gap and provide a comprehensive view of the state of the art in the national context.

3. Methodology

We carried out a systematic mapping study using the methodology presented by [Sinoara et al. 2017] and [Pelissari et al. 2022]. This choice was because they are based on the methodological path proposed by the seminal work [Kitchenham et al. 2007]. Systematic mapping is a bibliographic review technique that, although it differs from a systematic review due to the depth and breadth of the analyzed studies, follows a well-defined protocol and can be used to obtain a mapping of publications on some subject or field, identifying research gaps and areas that require the development of primary studies [Sinoara et al. 2017]. Following, the three phases are described to know: planning, conducting, and reporting.

3.1. Planning

In the planning phase, the protocol was defined, in which the research questions are described, as well as the research process with the sources in which the studies were mapped, and the studies selection guided by the inclusion and exclusion criteria.

Research Questions (RQ). The research problem tackled is related to a lack of information about methods and techniques for analyzing social media in Brazil. To solve this knowledge gap, a series of research questions were formulated that sought to be answered through systematic mapping. The RQs are presented below:

- RQ1: Which NLP tools and techniques are primarily used in scientific events under scrutiny?
- RQ2: What are the sources and nature of the data used in social media analysis?
- RQ3: What are the most used evaluation metrics in studies using NLP for social media analysis?

Search Process. The research process consisted of a manual search of the scientific events' proceedings, such as conferences, symposiums, meetings, and workshops between 2020 and 2022, listed in Table 1. These works are available in two digital library research sources: the SBC-OpenLib (SOL) of the Brazilian Computer Society (SBC) and SpringerLink. Therefore, the choice of these events is justified because they are considered important bases of national research in studies related to the areas of NLP, AI, and Computational Intelligence, as pointed out by [Lobato et al. 2021], [Carvalho et al. 2022] and [Pardo et al. 2010].

Study Selection. Inclusion and exclusion criteria were applied to select the most relevant works on NLP techniques applied to social media analysis. The inclusion and exclusion criteria are given in Table 2. Inclusion criteria embrace papers in the events'

Table 1. List of the relevant events considered in this mapping.

Source	Acronym	Edition
Brazilian Conference on Intelligent Systems	BRACIS	2020-2022
Brazilian Workshop on Social Network Analysis and Mining	BraSNAM	2020-2022
Encontro Nacional de Inteligência Artificial e Computacional	ENIAC	2020-2022
International Conference on Computational Processing of the Portuguese Language	PROPOR	2020-2022
Symposium in Information and Human Language Technology	STIL	2021

proceedings that address techniques, models, text mining tools, and social media analysis. Subsequently, scientific papers outside the inclusion criteria, such as publications in languages other than Portuguese or English, works irrelevant to the NLP, and social media analysis, were removed from our study. This selection followed the order (i) title, abstract, and keywords; (ii) introduction and conclusion and (iii) full paper. Papers that addressed systematic mapping or systematic literature review were also excluded.

Table 2. Inclusion (IC) and Exclusion (EC) Criteria for Selecting Relevant Studies.

Inclusion Criteria (IC)
IC1: Papers that address techniques, models, text mining tools, and textual analysis in social media analysis.
Exclusion Criteria (EC)
EC1: Papers that are outside the inclusion criteria
EC2: Publications written in languages other than Portuguese or English.
EC3: Papers that are not relevant to NLP and social media analysis based on the title, abstract, keywords, introduction, and conclusion.
EC4: Papers of systematic mapping or systematic literature review.

3.2. Conduction

As mentioned before, we evaluated papers published between 2020 and 2022 from BRACIS, BraSNAM, ENIAC, PROPOR, and STIL, which resulted in 654 papers. Two authors read the papers individually and assessed whether the works met the inclusion and exclusion criteria in Table 2. After rigorously applying these criteria, 468 papers were excluded. The exclusion occurs because many of these studies involved manipulating multimedia data, including images, videos, and audio, or creating, describing, or annotation corpora; thus, they are not within the scope of this systematic mapping, which aims to evaluate studies that address text mining or text analysis. Based on the established inclusion criteria, 186 papers were selected for data extraction. Fig. 1 presents the number of works included and excluded in each selected event.

Data Extraction. For extracting data related to research questions, we created a spreadsheet using Google Sheets to organize the attributes. The worksheet covers the following attributes: paper metadata (title, year, authors *etc*); data source (*e.g.*, Twitter, Reddit, and Facebook) and data nature (*e.g.*, corpus built/collected, already available or not described) responding to RQ2; Tool and Technology (*e.g.*, NLTK and spaCy); tasks (*e.g.*, preprocessing, text classification, and sentiment analysis); and techniques (*e.g.*, TF-IDF and BERT) related to RQ1. Additionally, evaluation measures were identified (*e.g.*, F1-Score, and

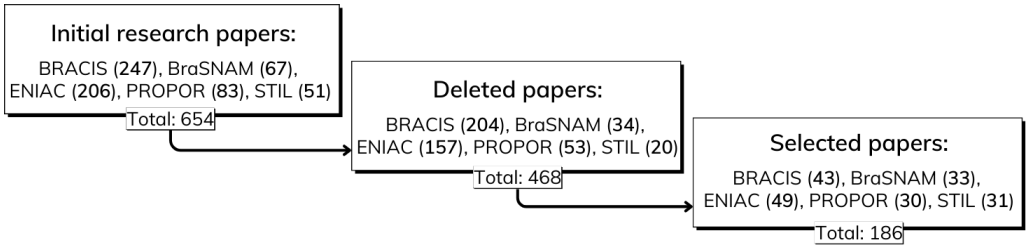

Figure 1. Overview of the systematic mapping protocol application.

cosine similarity) are used to respond to RQ3, and the development environment (*e.g.*, Python and Jupyter Notebook) was also included. In summary, during the complete paper reading process, these attributes were identified and indexed in the spreadsheet according to the order in which they were read. This information was mainly extracted from the materials and methods described in the papers.

Data analysis. An inductive approach was adopted to extract and analyze information from the collected qualitative data. For this, an exploratory analysis was performed using the Python 3.9.13 programming language with the aid of the Jupyter Notebook interactive programming environment. We used Pandas for data manipulation and Matplotlib and Seaborn for data visualization, following the other reviews such as [Lequertier et al. 2021, Almeida et al. 2020]. It is important to note that the attributes were pre-processed to remove unnecessary spaces and accents and convert the strings to lowercase. Therefore, qualitative data were extracted by developing a dictionary for counting words. Worth mentioning that all material produced during the conduction of this systematic mapping is publicly available in a repository on GitHub[1]. These papers will be described and explored in the results section.

4. Results and Discussion

In this section we answer and discuss the RQ. With this, a general and exploratory analysis of the works is presented, approaching the most frequent NLP tools and techniques in social media analysis, followed by the sources and nature of the data used in these analyses. Finally, we present the evaluation metrics identified in the studies.

4.1. Exploratory analysis

A total of 186 papers were selected from the 654 works published in the above-mentioned conferences, considering a time frame from 2020 to 2022. It means that 30% (186) papers are addressing NLP applied to social media analysis. The temporal distribution by year of publication is shown in Fig. 2.

Fig. 2 shows a significant increase in publications in 2021. This trend can be attributed to the growing amount of content generated on social networks due to the COVID-19 pandemic [Pachucki et al. 2022, Rosen et al. 2022]. It is essential to highlight that some events, such as STIL and PROPOR, have a biennial periodicity. Also, while conducting this analysis, the diffusion of works in other areas of AI and Computational Intelligence was noticed in the event proceedings, such as BRACIS and ENIAC in the year 2022. This trend of exploration and exploitation of new research fields may

[1]https://github.com/fabiolobato/ENIAC23-SysMapping

Figure 2. Distribution of publications by year and events.

have impacted the proportion of specific textual analysis studies within the scope of this mapping.

Although BRACIS and ENIAC have more published papers (gross), considering the scope of this systematic mapping, BraSNAM, PROPOR, and STIL present a higher percentage of works included in our analysis. For instance, STIL presents 61% and BraSNAM 49% of papers included. It means that they are applying NLP to social media analysis in-depth, while BRACIS and ENIAC have a broader scope. These results corroborate the conclusions discussed by [Souza et al. 2018].

4.2. Tools and Techniques in NLP for social media analysis

In this section, we perform an exploratory analysis in order to answer the research questions. For this, we conducted a quantitative analysis considering the count of each unique attribute extracted in each selected study. The same applies to the following sections.

- RQ1: Which NLP tools and techniques are primarily used in scientific events under scrutiny?

Considering the 186 analyzed studies, we identified 135 tools assigned to NLP tasks, corresponding to more than 81% of the works. Fig. 3 presents the ten most frequent tools. Scikit-Learn is the most prevalent, corresponding to 37% of mentions. This tool was specifically developed for practical ML applications and can be used in various stages of the NLP pipeline. It is widely used in text classification tasks, in studies such as [Vitório et al. 2022] and [Cordeiro et al. 2022], feature extraction through topic modeling analysis in [de Sousa et al. 2020], and resources for sentence representations as in [Aragy et al. 2021]. Thus, this tool stands out for its versatility, allowing researchers to explore different approaches and models for their text analysis.

For text preprocessing tasks, NLTK is the most used tool, applied in 39 studies (26%), being in second place in the overall ranking. NLTK is an open-source platform for NLP that supports several tasks, such as tokenization, lemmatization, stopwords removal, POS tagging, *etc*. Followed by spaCy, mentioned by 23 papers (15%), a Python library that provides features for tasks like POS tagging, NER, parsing, text classification, and

stemming. Fig. 3 also reveals a large number of underutilized tools, such as TensorFlow and PyTorch, which are ML and DL frameworks for building and deploying language models. These tools play a crucial role in data processing and analysis and are in line with the techniques that will be described later.

Besides, it is crucial to point out that 36 studies (19%) did not clearly describe the tools used. It means that the experiment's replicability is impaired/unfeasible. Apart from the efforts of the Brazilian community to stick to open science [Laender et al. 2020], this research-find raises a red flag to Technical Programs Committees to include this aspect in their evaluation systems.

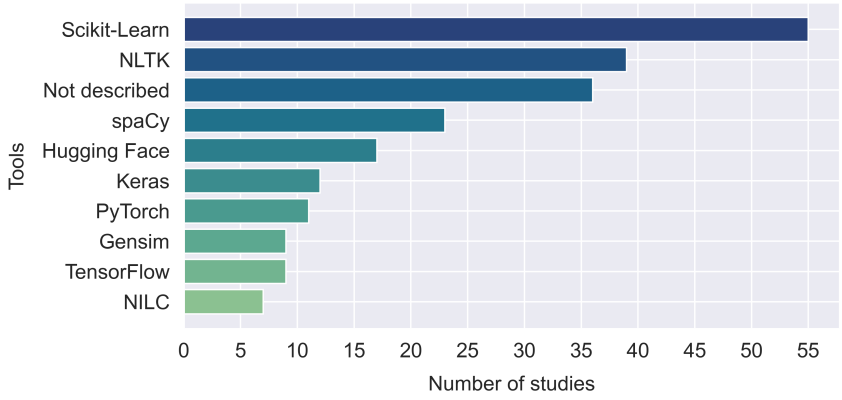

Figure 3. The 10 most frequent tools in social media analysis.

A total of 275 techniques were identified, where only 3 of the studies did not clearly describe the technique used. Fig. 4 shows that the most frequent technique is BERT, used by 63 studies (34%), which consists of a language model based on Transformers that stood out for its performance in NLP tasks, mainly for pre-training representations of unlabeled texts [Devlin et al. 2018]. In particular, the high usage of the pre-trained BERT model indicates the recognition of its effectiveness in capturing semantic relationships and its applicability in several tasks such as natural language inference [Nanclarez et al. 2022], text classification [Serras and Finger 2021, Ferraz et al. 2021], sentiment analysis [Britto et al. 2022] and, entities extraction through the tokens classification in [Lochter et al. 2020]. Besides, some BERT variants are also present in our analysis, worth mentioning BERTimbau, discussed in 32 papers (17%); Multilingual BERT (M-BERT) used in 15 studies (8%); and BERTopic mentioned in 6 works (3%).

Regarding the techniques, 105 of the studies reported splitting data into training and testing sets. Based on this, many DL and ML models were applied, in such a way that there is the presence of several other DL techniques, such as Word embedding applied in 53 studies (29%), Word2Vec in 29 papers (16%), LSTM used in 29 studies (16%), Multilayer Perceptron (MLP) in 20 papers (11%), GloVe present in 20 studies (11%), FastText in 16 studies (9%), Convolutional Neural Network (CNN) mentioned in 16 works (9%) and Bidirectional LSTM with Conditional Random Fields (BiLSTM-CRF) in 16 studies (9%).

From the ML techniques identified, 48 studies (26%) mention SVM, widely used for classifying and training labeled textual data. SVM is followed by Naïve Bayes, used in

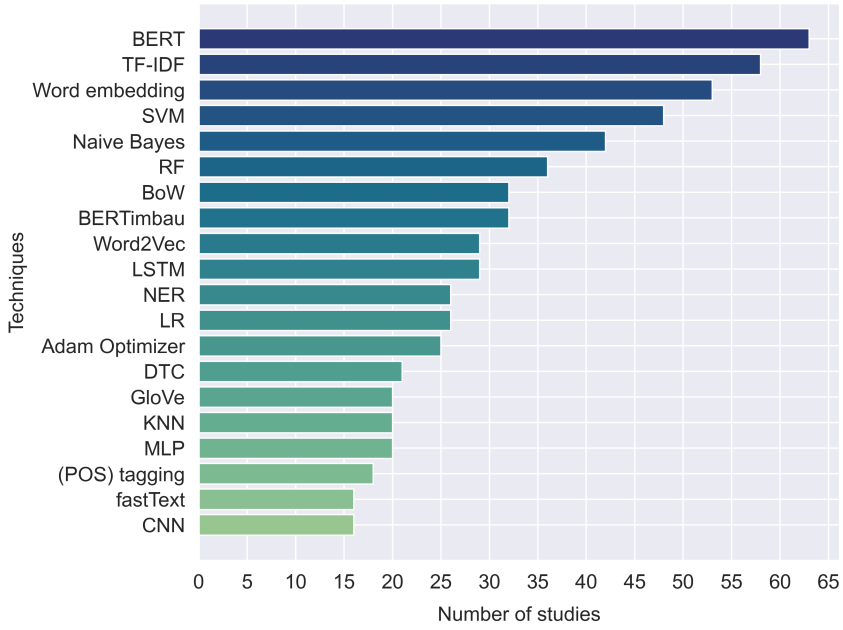

Figure 4. The 20 most frequent techniques in social media analysis.

42 studies (23%); Random Forest (RF) present in 36 works (20%); LR in 26 papers (14%); and Decision Tree Classifier (DTC) in 21 studies (12%). Other language models and text representations are also present. When interpreting the results, it is observed that the techniques most mentioned in the studies reflect current trends in the NLP Research field, as stated by [Khurana et al. 2023]. In addition, a comparison of the ten most used tools between events was performed, as shown in Fig. 5. This Figure helps identify researchers' preferences regarding NLP techniques in text analysis.

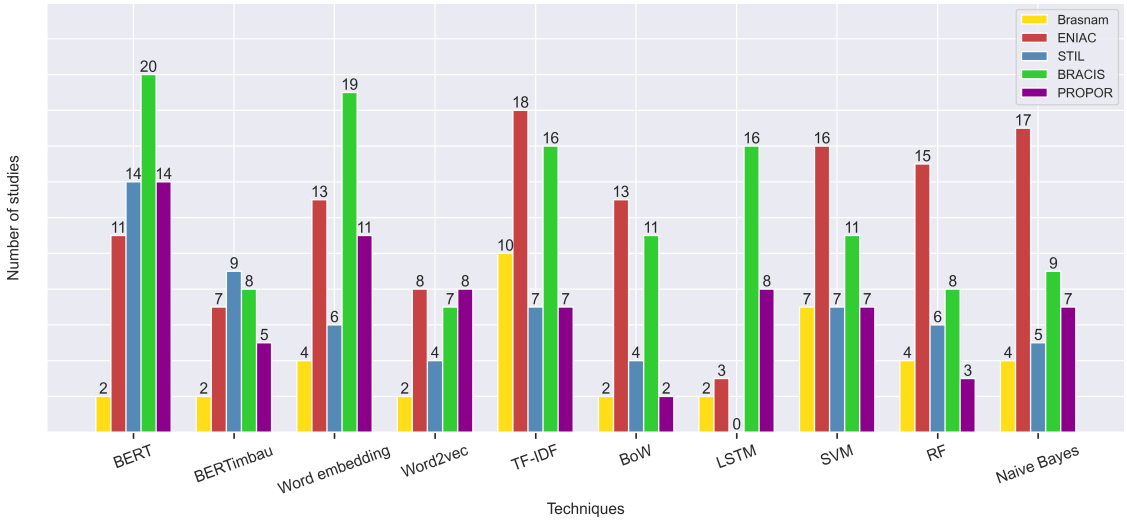

Figure 5. Comparison of the 10 most frequent techniques in social media analysis by event.

As shown in Figure 5, BRACIS has the highest prevalence of works using DL techniques, mainly for text classification, with language models such as BERT, Word

embedding, and LSTM. These results suggest a significant interest in using approaches based on pre-trained language models, distributed word representations, and recurrent neural networks in social media analysis - this research-find is in consonance with [Gururangan et al. 2020]. In contrast, ENIAC addresses more works that use text classification and text representation models in their analyses, such as SVM, RF, Naïve Bayes, and TF-IDF. These results show the use of traditional ML techniques in social media analysis in studies published in ENIAC. It is also observed that the BERTimbau pre-trained model has been widely used by papers published in STIL, BRACIS, ENIAC, and PROPOR. This model emerged as an approach proposed by [Souza et al. 2020] to deal with NLP tasks [Hammes and de Freitas 2021, Cortiz et al. 2021].

4.3. Data Sources

Regarding research question 2:

- RQ2: What are the sources and nature of the data used in social media analysis?

The most frequent data sources in the studies were identified during the papers' information extraction. By analyzing the word cloud in Fig. 6, it is possible to observe that Twitter is the most frequent data source, indicating that researchers widely explore this platform. In addition to Twitter, other sources, such as Google News, are highlighted. It is worth mentioning that other datasets are derived from Google News, such as the *Avaliação de Similaridade Semântica e Inferência Textual*[2] (ASSIN) which consists of a set of data built with 10,000 sentence pairs of real journalistic texts. Some other data sources are also observed in Fig. 6, a non-exhaustive list includes: Wikipedia, UOL[3], TripAdvisor, G1[4], *Folha de São Paulo*, Amazon, Judiciary System platforms, and e-commerce platforms.

Figure 6. Word cloud of data sources identified by our analysis.

From an overall analysis of the results presented in Fig. 6, we can assert that the data sources identified represent different contexts, ranging from user opinions,

[2]http://nilc.icmc.usp.br/assin/
[3]https://www.uol.com.br/
[4]https://g1.globo.com/

news, and product information in e-commerce to legal data. These research-findings are consistent with [Choi et al. 2020], [Zachlod et al. 2022], [Souza et al. 2018], and [Júnior et al. 2020].

Twitter, in particular, receives much attention from scholars due to its instantaneous nature and the amount of textual data available. Regarding the news datasets, as pointed out by [Souza et al. 2018], many of the studies investigate fake news detection, also proposing news datasets - such as the ones focused on the Portuguese language [Batista Filho et al. 2021, Charles et al. 2022]. One of the reasons for including these data sources in the studies reflects the search for more effective solutions to combat fake news and contributes to the development of specific resources and algorithms.

Additionally, how the data were obtained was also investigated. Considering the 186 studies analyzed, 110 works used corpus collected or "self-made", which involves creating or collecting specific data for the work. On the other hand, 75 works used already available corpora, that is, data sets readily available for analysis. Only one work did not describe the nature of the used corpus.

4.4. Evaluation measures

In this subsection, we explore the metrics used to evaluate and validate the data in the studies, aiming to answer the following research question:

- RQ3: What are the most used evaluation metrics in studies using NLP for social media analysis?

There was evidence of great variety in the evaluation measures/strategies employed. For instance, several metrics were used to evaluate the performance of NLP classification models. Among the 101 metrics identified, the most used was the F1-Score, present in 108 studies (58%). This metric represents the harmonic mean between Recall and Precision, providing a balanced measure of model performance. The Recall was present in 81 (43%) and Precision in 78 (42%) papers, followed by Accuracy, mentioned by 55 works (30%).

This reveals a common approach to validating text classification models since many studies employ evaluation metrics to compare the performance of different DL and ML models as in [Cordeiro et al. 2022, Gumiel et al. 2021]. This analysis corroborates the findings of [Souza et al. 2018]. In addition, the prevalence of F1-Score possibly is related to imbalanced datasets, a common phenomenon in text classification. In addition to these evaluation measures, researchers also use different approaches in generating dataset splits for testing and validation, including the Cross-Validation method cited in 59 (32%) papers. Also called K-fold cross-validation, or simply k-fold, it involves the random division of the database into K subsets of similar size, where K is previously defined [Berrar 2019].

5. Conclusions

In this paper, we discussed a systematic mapping study to analyze NLP techniques used in social media analysis in order to investigate and understand current trends and challenges in this prominent research field. We identified the main scientific events in the area (BRACIS, BRaSNAM, ENIAC, STIL, and PROPOR) and analyzed 654 papers published

between 2020 and 2022, selecting 186 relevant works. This period of time was chosen given the dynamic nature of the area, hence its rapid obsolescence. In short, this systematic mapping reveals a dynamic and constantly evolving scenario in the field of NLP to social media analysis. We shed light on the communities/events' evolution and perspectives; the technologies (tools) and techniques used (RQ1); the data sources (RQ2), and evaluation measures (RQ3).

Besides, we critically evaluate the research applicability, considering the source-code availability. In summary, we contribute to the body of knowledge by providing an overview of NLP tools and techniques used in social media analytics, also mapping data sources, and evaluation measures. It is essential to highlight that we are following Open Science Principles by providing all our data in a publicly available GitHub repository, allowing the research reproducibility. This work can be helpful for researchers and practitioners interested in exploring the potential of these tools and techniques, having a clear picture of research gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media.

Worth mentioning some threats to the study's validity. One is the limitation of the selected events, which may only partially represent some research areas in social media analysis. Another threat is the possibility of publication bias; since the selected papers are those available in the proceedings of specific events. For future work, we plan to expand our data sources to relevant international conferences (*e.g.*, AAAI Conference on Artificial Intelligence, International AAAI Conference on Web and Social Media, *etc*), aiming to compare the current status of Brazilian research programs with the international ones. Finally, we plan to expand our research questions to verify if the tools/techniques mapped are sufficient and their limitations.

Acknowledgments

The authors acknowledge the support of the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) - DT - 308334/2020, PIBIC - 132759/2022-5, PIBITI - 141598/2022-0; the *Fundação Amazônia de Amparo a Estudos e Pesquisas* (FAPESPA) - PRONEM-FAPESPA/CNPq nº 045/2021; and by *Acordo de Cooperação Técnica* N°02/2021 (Processo N°38328/2020 -TJ/MA).

References

Aichner, T., Grünfelder, M., Maurer, O., and Jegeni, D. (2021). Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Almeida, G. R., Guimarães, I., Jacob Jr, A. F., and Lobato, F. M. (2020). Fontes de dados gerados por usuários: quais plataformas considerar? In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 25–36. SBC.

Appel, G., Grewal, L., Hadi, R., and Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing science*, 48(1):79–95.

Aragy, R., Fernandes, E. R., and Caceres, E. N. (2021). Rhetorical role identification for portuguese legal documents. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Proceedings, Part II 10*, pages 557–571. Springer.

Balaji, T., Annavarapu, C. S. R., and Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395.

Batista Filho, A. P., da Conceiçao Araújo, D., Ferreira, M. A. D., and de Mattos Neto, P. S. G. (2021). Fake news detection about covid-19 in the portuguese language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*.

Berrar, D. (2019). Cross-validation. In Ranganathan, S., Gribskov, M., Nakai, K., and Schönbach, C., editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Academic Press, Oxford.

Britto, L. F., Pessoa, L. A., and Agostinho, S. C. (2022). Cross-domain sentiment analysis in portuguese using bert. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 61–72. SBC.

Carvalho, L. P., Murakami, L., Suzano, J. A., Oliveira, J., Revoredo, K., and Santoro, F. M. (2022). Ethics: What is the research scenario in the brazilian conference bracis? In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 624–635. SBC.

Charles, A. C., Ruback, L., and Oliveira, J. (2022). Fakepedia corpus: A flexible fake news corpus in portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022*, pages 37–45. Springer.

Choi, J., Yoon, J., Chung, J., Coh, B.-Y., and Lee, J.-M. (2020). Social media analytics and business intelligence research: A systematic review. *Information Processing & Management*, 57(6):102279.

Cordeiro, F., Rabelo, R. d. A. L., and Moura, R. S. (2022). Classification of irregularity communications in public ombudsmen using supervised learning algorithms. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 704–715. SBC.

Cortiz, D., Silva, J. O., Calegari, N., Freitas, A. L., Soares, A. A., Botelho, C., Rêgo, G. G., Sampaio, W., and Boggio, P. S. (2021). A weakly supervised dataset of fine-grained emotions in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 73–81. SBC.

de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., and Mattos, D. M. (2021). Identifying fake news on social networks based on natural language processing: trends and challenges. *Information*, 12(1):38.

de Sousa, G. N., Guimaraes, I., Jacob Jr, A. F., and Lobato, F. M. (2020). Análise comparativa das principais plataformas de reclamações online: implicações para análise de mídia social em negócios. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 154–165. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Ferraz, T. P., Alcoforado, A., Bustos, E., Oliveira, A., Gerber, R., Müller, N., d'Almeida, A. C., Veloso, B., and Costa, A. R. (2021). Debacer: a method for slicing moderated debates. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 667–678. Sociedade Brasileira de Computação-SBC.

Gumiel, Y. B., Lee, I., Soares, T. A., Ferreira, T. C., and Pagano, A. (2021). Sentiment analysis in portuguese texts from online health community forums: data, model and evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 64–72. SBC.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Hammes, L. O. A. and de Freitas, L. A. (2021). Utilizando bertimbau para a classificação de emoções em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 56–63. SBC.

Hassani, A. and Mosconi, E. (2022). Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing smes. *Technological Forecasting and Social Change*, 175:121416.

He, W., Zhang, W., Tian, X., Tao, R., and Akula, V. (2019). Identifying customer knowledge on social media through data analytics. *Journal of Enterprise Information Management*, 32(1):152–169.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hou, Q., Han, M., and Cai, Z. (2020). Survey on data analysis in social media: A practical application aspect. *Big Data Mining and Analytics*, 3(4):259–279.

Júnior, E. G. S. L., de Sousa, G. N., Junior, A. F. L. J., and Lobato, F. M. F. (2020). Ferramentas para análise de mídias sociais: Um levantamento sistemático. *Anais do Computer on the Beach*, 11(1):389–396.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*.

Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Laender, A. H. F., Medeiros, C. M. B., Cendes, I. L., Barreto, M. L., Van Sluys, M.-A., Almeida, U. B. d., et al. (2020). Abertura e gestão de dados: desafios para a ciência brasileira. Technical report, Academia Brasileira de Ciências.

Lequertier, V., Wang, T., Fondrevelle, J., Augusto, V., and Duclos, A. (2021). Hospital length of stay prediction methods: a systematic review. *Medical Care*.

Lobato, F. M., de Sousa, G. C., and Jacob Jr, A. F. (2021). Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 217–228. SBC.

Lochter, J. V., Silva, R. M., and Almeida, T. A. (2020). Deep learning models for representing out-of-vocabulary words. In *Intelligent Systems: 9th Brazilian Conference, BRACIS, Proceedings, Part I*, pages 418–434. Springer.

Mirzaalian, F. and Halpenny, E. (2019). Social media analytics in hospitality and tourism: A systematic literature review and future trends. *Journal of Hospitality and Tourism Technology*, 10(4):764–790.

Nanclarez, R. G., Roman, N. T., and da Silva, F. J. (2022). Generalizing over data sets: a preliminary study with bert for natural language inference. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 602–611. SBC.

Pachucki, C., Grohs, R., and Scholl-Grissemann, U. (2022). Is nothing like before? covid-19–evoked changes to tourism destination social media communication. *Journal of Destination Marketing & Management*, 23:100692.

Pardo, T., Gasperin, C., de Medeiros Caseli, H., and Nunes, M. d. G. V. (2010). Computational linguistics in brazil: an overview. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*.

Pedroso, P. M., Lobato, F. M., de JV Sá, E., and Jacob, A. F. (2022). Handling out of vocabulary words at the semantical level using recurrent neural networks. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 88–94. IEEE.

Pelissari, R., Alencar, P. S., Amor, S. B., and Duarte, L. T. (2022). The use of multiple criteria decision aiding methods in recommender systems: A literature review. In *Brazilian Conference on Intelligent Systems*, pages 535–549. Springer.

Rosen, A. O., Holmes, A. L., Balluerka, N., Hidalgo, M. D., Gorostiaga, A., Gómez-Benito, J., and Huedo-Medina, T. B. (2022). Is social media a new type of social support? social media use in spain during the covid-19 pandemic: A mixed methods study. *International Journal of Environmental Research and Public Health*, 19(7):3952.

Serras, F. R. and Finger, M. (2021). verbert: Automating brazilian case law document multi-label categorization using bert. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 237–246. SBC.

Sinoara, R. A., Antunes, J., and Rezende, S. O. (2017). Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society*, 23:1–20.

Souza, E., Costa, D., Castro, D. W., Vitório, D., Teles, I., Almeida, R., Alves, T., Oliveira, A. L., and Gusmão, C. (2018). Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference*. Springer.

Vitório, D., Albuquerque, H. O., Souza, E. P. R., Oliveira, A. L. I. d., Barros, F., and Prudêncio, R. B. (2022). Análise do posicionamento dos usuários do twitter acerca da vacinaçao infantil contra a covid-19 no brasil. *Anais*.

Zachlod, C., Samuel, O., Ochsner, A., and Werthmüller, S. (2022). Analytics of social media data–state of characteristics and application. *Journal of Business Research*.

Zhang, C. and Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23:100224.