B2T: A Dataset of Tweets in Portuguese Language about Brazilian Banks

Gabriel K. Kakimoto [1,2,3], Seyed J. Haddadi [1,2,3], Patrick M. Araújo [1,2], Fillipe S. Silva [1,2,3], Julio C. dos Reis [1,2], Marcelo S. Reis [1,2,3]

[1]Instituto de Computação, Universidade Estadual de Campinas (UNICAMP);

[2]Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC);

[3]Artificial Intelligence Laboratory (Recod.ai)

g234878@dac.unicamp.br

{seyed, fillipesantos, p217144, dosreis, msreis}@unicamp.br

Abstract. *Sentiment Analysis models have numerous applications, including evaluating business value from customers' perspectives through comments and reviews. This capability helps businesses understand the public perceptions about their products and services and identify areas for improvement. A significant limitation in developing such models for the Portuguese language is the lack of labeled datasets, which restricts practical model training. This article addresses this issue by collecting 375,912 comments from Twitter/X, focusing specifically on comments about Brazilian banks due to the public's widespread use of their services. The labeled dataset currently contains 1,096 comments labeled as Positive, Neutral, or Negative. We present results in fine-tuning Sentiment Analysis models based on this dataset. We found it holds the potential to provide insights into customer perceptions and market trends within the banking sector. By leveraging this dataset, businesses can gain a valuable understanding of their market position and areas for service improvement.*

1. Introduction

Sentiment analysis, the computational study of opinions, sentiments, and emotions expressed in texts [Liu et al. 2010], aims to determine whether a natural language text conveys a positive, negative, or neutral sentiment. This field has gained significant attention and has been applied across various domains, including business popularity analysis, customer feedback analysis, and market research [Drus and Khalid 2019]. A notable application of this technique is in evaluating business performance. For instance, Saragih et al. [Saragih and Girsang 2017] analyzed comments from Facebook and Twitter/X to assess the performance of online transportation services.

Most sentiment analysis models rely on supervised learning, which requires labeled data to perform effectively. However, labeled datasets in the Portuguese language are scarce. This scarcity is due to several factors, including limited resources and initiatives to create and maintain such datasets, the substantial labor required to create and label large amounts of data, and the comparatively smaller volume of texts in Portuguese versus languages like English. This shortage of datasets in Portuguese significantly hinders the development of high-quality models for this language [Pereira 2021].

To address this issue and inspired by the application of sentiment analysis in evaluating business performance, we decided to create a dataset comprising comments from Twitter/X about Brazilian banks, each labeled with the sentiment expressed. We chose Twitter/X for this research due to its unique characteristics: it is one of the most widely used social networks, predominantly composed of text data with a limit of 280 characters per post. Additionally, it hosts a wide array of comments on various subjects. Given our aim to analyze data that can benefit companies, we decided to focus on comments related to Brazilian banks. Due to their popularity and the significance of their services, these banks generate a substantial number of comments. The banks included in our analysis are Itaú, Bradesco, Santander, Banco do Brasil, BTG Pactual, Nubank, Inter, Banco Pan, Sofisa Direto, and BRB Oficial. At the time the data was crawled, banks such as Caixa Econômica Federal were not included. However, because of recent analysis, this bank will be considered in future works.

To evaluate the dataset, we analyzed the agreement between the raters' labels using the Cohen Kappa metric, viewing it as an indication of the reliability and quality of the labels. Additionally, the dataset was used to train three traditional sentiment analysis models (XGBoost, SVM, and Logistic Regression) and a variant of the BERT model tailored for Portuguese texts: BERTimbau [Souza et al. 2020]. The Cohen Kappa showed that most of the raters have a moderate agreement with each other, which could be improved by better instructing them on how to properly label the data and improving the methodology of the labeling process.

The availability of datasets focusing on sentiment analysis in the banking sector is limited, especially in the Portuguese language. Among the notable datasets, the Post Bank Customer Review [Plotnikov et al. 2020] and Phrase Banking [Malo et al. 2014] stand out, containing reviews from a Russian finance website and curated articles from financial and economic sources, respectively. However, both datasets lack Portuguese language content.

The primary contribution of this study lies in the collection and annotation of a comprehensive corpus of comments in the Portuguese language about Brazilian banks. This dataset is specifically designed to address the current shortage of labeled corpora for sentiment analysis in Portuguese within the banking sector. Beyond its utility in training sentiment analysis models, this corpus holds the potential for generating valuable business insights and facilitating experimental research on the application of such data in the banking industry. The dataset is available in the Github Repository: https://github.com/GabrielKakimoto/B2T-A-Brazilian-Tweet-Banks-Dataset-in-Portuguese-Language

The remaining of this article is organized as follows: Section 2 reviews related work. Section 3 outlines our methodology with the steps and tools used to collect and label the dataset. Section 4 presents a detailed description of the dataset by including results of the effectiveness of the sentiment analysis models using the dataset. Section 5 discusses the reached findings and implications. Finally, Section 6 wraps up the concluding remarks.

2. Related Work

Due to the large quantity of tweets generated every day and the easy access to them, other works also used this social media to collect textual data. In the dataset TweetSentBR [Brum and das Graças Volpe Nunes 2017], the authors combined the Twitter/X API with web crawlers to collect 15.000 Tweets in Portuguese that talked about TV shows. Following the recommendations from [Hovy and Lavid 2010], the tweets were labeled as Positive, Negative, and Neutral by different volunteers. The final published dataset was composed of the comment's id, the hashtag used to search and select the comment, each label selected by the volunteers, how many volunteers were in doubt when labeling the comment, the sentiment expressed in the comment, and an indication regarding if the comment is in the test or train dataset. An aspect of the final dataset is its unbalance, the majority of the comments are Positive, which represents approximately 45% of the dataset, while the Negative and Positive classes represent 25% and 29%, respectively.

On the other hand, some datasets utilize the available metadata in the comments to infer the expressed sentiment. For instance, in the UTLCorpus dataset [Sousa et al. 2019], data was extracted from a Brazilian social network for movies and reviews from the Google Play App Store. In this dataset, each comment was accompanied by a score given to the movie or product and the number of "likes" the review received. The authors used the score to classify the sentiment of the review as Positive or Negative, while the number of "likes" was used to determine the review's relevance, which is a novelty approach of this work.

A more recent work [Alves et al. 2024] created a database composed of comments from movie trailers from the Netflix YouTube channel. The authors collected 2.496 comments that, like in the previously mentioned works, were labeled as Positive, Negative, and Neutral. A unique aspect of this work is the method used in the labeling process. In this research, the authors used the GPT3.5 from the Chat GPT User Interface. This work, as well as the two previous works, mention, in their motivation, the lack of labeled corpus for sentiment analysis in Portuguese, and aimed to address this issue.

Another interesting work [França et al. 2017], collected comments from Twitter/X that talked about the Brazilian Uprising that occurred between June and July of 2013. The collected comments were labeled by 3 different volunteers, and they were classified as Positive, Neutral and Negative. The classification analyzed if the comment showed "agreement (positive), disagreement (negative) or neither (neutral) to the protests". The dataset showed usefulness in the analysis of the protests of 2013 and in the evaluation of machine learning models for the processing of texts in Portuguese Language.

While previous studies have addressed the issue of the scarcity of datasets in the Portuguese language, none have specifically examined their application in the banking sector. In the article "Sentiment Analysis on Banking Feedback and News Data using Synonyms and Antonyms" [Mohanty and Cherukuri 2023], the author investigates the Post Bank Customer Review [Plotnikov et al. 2020] and Phrase Banking [Malo et al. 2014] datasets. Although these datasets are not in Portuguese, they contain labeled texts pertinent to the banking sector. Both datasets are unbalanced and categorized into three classes: Positive, Negative, and Neutral. This work exemplifies the utility of datasets focused on customer feedback in the banking industry. It underscores the applica-

tion of sentiment analysis in customer feedback, demonstrating how it can yield insights for brand monitoring, product and service evaluation, fraud detection, market research, and competitor analysis.

This study introduces a novel dataset consisting of labeled comments categorized into three classes, specifically focused on the banking sector. These comments, sourced from Twitter/X, are particularly valuable for conducting sentiment analysis tasks within the Portuguese banking domain.

3. Methodology

This section delineates the comprehensive methodology employed for data collection, storage, and labeling of the comments and their application in training four sentiment analysis models. Figure 1 presents our defined methodology. The first part of the methodology (cf. Subsection 3.1) focuses on creating the dataset. This includes the data collection process, storage methods, and labeling procedures. The second part (cf. Subsection 3.2) focuses on applying the dataset in training sentiment analysis models.

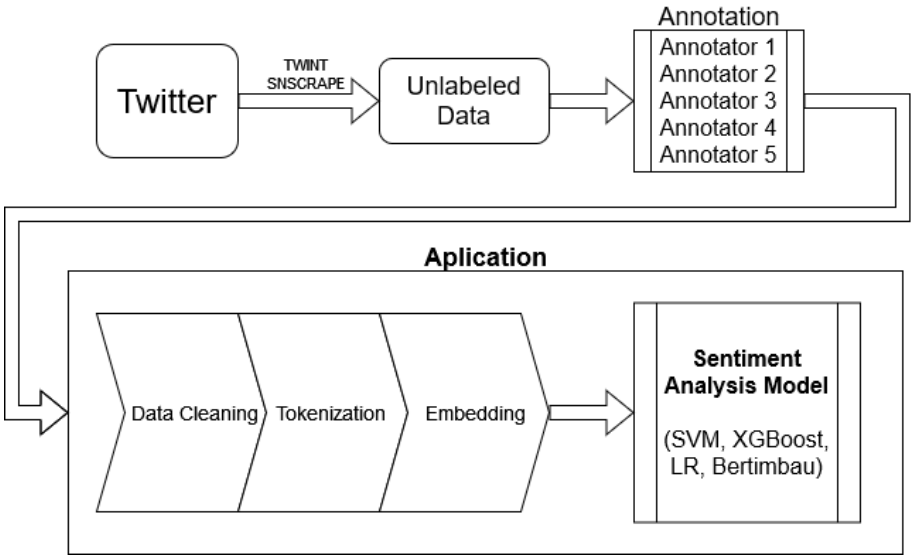


Figure 1. Methodology designed to create the dataset and train the sentiment analysis models

3.1. Dataset Creation

To create the dataset containing the comments from Twitter/X that talked exclusively about Brazilian banks, it was necessary to make a query to search comments talking about them without collecting retweets (reposted comments). The query used for this case was: ”[name of the bank] -is:retweet”. For example, for the bank Santander, it would be: ”Santander -is: retweet”. Then, using the scraping tools, such as TWINT [1] and SNScrape [2], we started collecting comments from Twitter/X using the query mentioned before. The data crawling started in September 2022 and stopped in June 2023. The crawled data

[1]https://github.com/twintproject/twint
[2]https://github.com/JustAnotherArchivist/snscrape

was stored in a "CSV" file containing the columns: Tweet ID (an anonymized and unique number that distinguishes a comment from another), text, date and time of posting, location (if available), language, and the name of the bank used in the query. A portion of the collected comments were randomly selected to be labeled by the volunteers. The comments were sent in batches of approximately 200 comments. After labeling, the comments would be put together and anonymized to create the label column in the labeled dataset. This column would contain a list of the labels of the raters.

The labeling process was conducted by five volunteers who independently categorized each comment without seeing each other's labels. The comments were classified as Positive (+1), Negative (-1), or Neutral (0). The decision to use three classes—Positive, Negative, and Neutral—aligns with standard practices in sentiment analysis, as seen in datasets like TweetSentBR, PostBank Customer Review, and Phrase Banking. This three-class approach also simplifies the labeling process and ensures clear distinctions between different sentiment categories. To ensure the understanding of the differences between each class, the following definitions were given to the raters:

- **Positive**: The text transmits a sentiment of pleasure, satisfaction, or compliment. The target of the compliment should be a Brazilian bank or something related to it (such as the bank's app, its product, or its services).
- **Neutral**: This label must be used in case the annotator cannot identify a sentiment (Positive or Negative). Factual sentences (example: Santander just opened a new bank in my city), uncertainty or doubt (example: I don't know if it's good or bad), and incomprehensible sentences. Tweets that were made by a robot or an automatically generated message should be labeled as neutral as well.
- **Negative**: The text transmits a sentiment of displeasure, complaint, disgust, or hate. The target of the compliment should be a Brazilian bank or something related to it (such as the bank's app, its product, or its services).

Each volunteer received a table with four columns (text, Negative, Neutral, and Positive), shown in Figure 2 (the text of the comments were censored in the picture to follow the Twitter/X guidelines on sharing data). The first column contained the tweet text, and the other three columns contained checkboxes for each sentiment. Volunteers were instructed to select only one sentiment per comment. Although some volunteers couldn't label all the data, we ensured that each tweet was labeled by at least three different people.

To evaluate the consistency and quality of these labels, we used the Cohen Kappa metric to measure inter-rater agreement. It's expected that the higher the agreement, the more accurate and reliable the dataset. Meanwhile, inconsistencies in the labeling might indicate that the sentiment expressed in the comment is difficult to define or a mislabeled by one of the raters.

3.2. Sentiment Analysis

To evaluate and measure the usability of the dataset in sentiment analysis models, the data was divided into train and test datasets, representing 0.8 and 0.2 of the original dataset. To use the data in the sentiment analysis models, the following steps were made:

Noise Removal: The process involves removing punctuation, special characters, emojis, numbers, and stopwords to enhance the performance of sentiment analysis and

Tweet_id	Text	Positive	Neutral	Negative
1645595520608477184	**********	☐	☐	☐
1621346107593822208	**********	☐	☐	☐
1614942777707737088	**********	☐	☐	☐
1664668826322059264	**********	☐	☐	☐

Figure 2. Snapshot of the table given for the annotation (following the Twitter/X guidelines on sharing data, the text were censored in the picture)

topic modeling models. Stopwords are common words like articles, prepositions, and pronouns (e.g., ”a,” ”the,” ”of,” ”to”) that contribute little informational value to the text. Eliminating these words reduces the number of words to be analyzed, speeding up model processing.

Tokenization: This step involves breaking the text into words and smaller pieces, known as tokens.

Word Embedding: This step is responsible for transforming text into numbers and vectors that can be processed by machine learning models. For the simple model (XGBoost, SVM, and Logistic Regression), the Bag of Words technique was applied to generate the word embeddings. In contrast, BERTimbau employed its own pre-trained WordPiece embedding model.

Model Training: the vectors resulting from the embedding process were used to train the sentiment analysis models. After completing the training, the test dataset was used to evaluate the accuracy, recall,l, and f1 major scores of the trained models.

4. Dataset B2T

4.1. Dataset Description

The labeled dataset consists of 1,096 comments randomly selected from an unlabeled dataset of tweets collected from Twitter/X. Each entry contains information such as the Tweet ID, posting date and time, user location (if available), language, the name of the bank used for the search query, and labels provided by volunteers. In compliance with Twitter/X’s data-sharing guidelines, the tweet text is not included in the published dataset. However, the comments can be retrieved by querying Twitter/X’s API using the provided Tweet IDs. Additionally, the data about the location regards the location of the account, which can be inputted by the user, instead of the geolocation of the tweet itself. Therefore, it is information the user provides and is not 100% accurate. Table 1 presents a dataset sample. The unlabeled data contains 375,912 tweets containing all the columns mentioned before except the labels from the raters. The only missing values in the labeled dataset are from the location column, with 1011 missing values. The dataset contains significantly fewer labeled data points than other published datasets like TweetSentBR. It is continually being updated with additional labels.

Upon analyzing the comments, it is evident that the number of tweets per bank varies. Banks like ”Nubank” and ”Itau” received significantly more tweets, as illustrated in Figure 3. Meanwhile, taking a closer look at the labeled data, it’s evident that it is unbalanced, with a majority of comments labeled as Neutral (0), followed by Negative (-1)

Table 1. Sample of the B2T Dataset

Tweet_id	Datetime	Location	Lang	Query	Labels
166678494 4645980161	2023-06-08 12:31:12+00:00	-	pt	itau	[0.0, 1.0, 1.0, 1.0, 1.0]
166675089 9937574912	2023-06-08 10:15:55+00:00	-	pt	itau	[1.0, 0.0, 1.0, 0.0]
166663391 7770637312	2023-06-08 02:31:05+00:00	Rio de Janeiro	pt	itau	[0.0, 0.0, 1.0, -1.0]

and Positive (1) classes (Figure 4a). Additionally, by measuring the Cohen kappa between the labels of each rater, we generated the heat map shown in Figure 4b. The Cohen kappa values range from 0.56 to 0.67. According to [McHugh 2012], values between 0.41 and 0.60 indicate moderate agreement, and values between 0.61 and 0.80 indicate substantial agreement. Therefore, the heat map shows that the raters have, on average, a moderate agreement in sentiment labels.

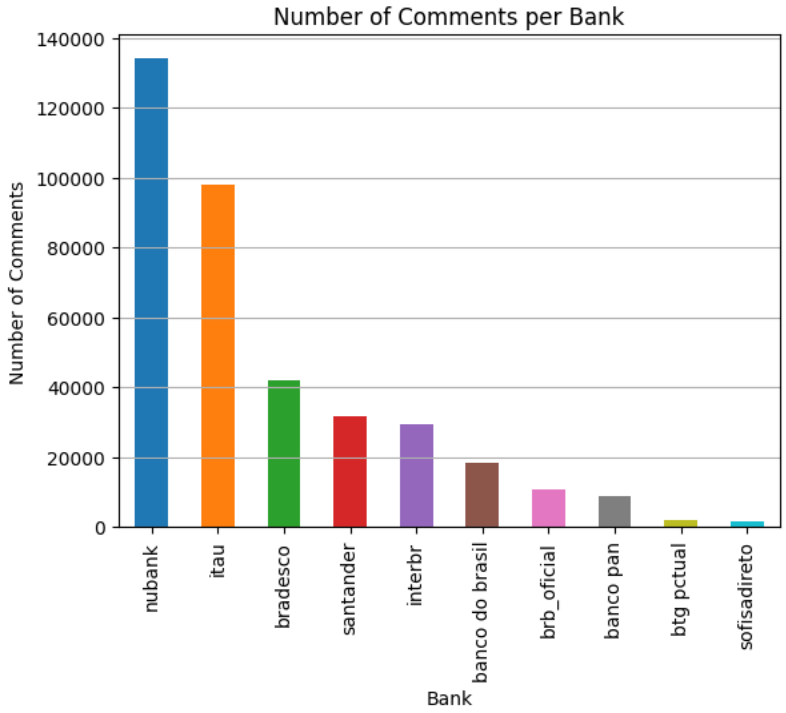
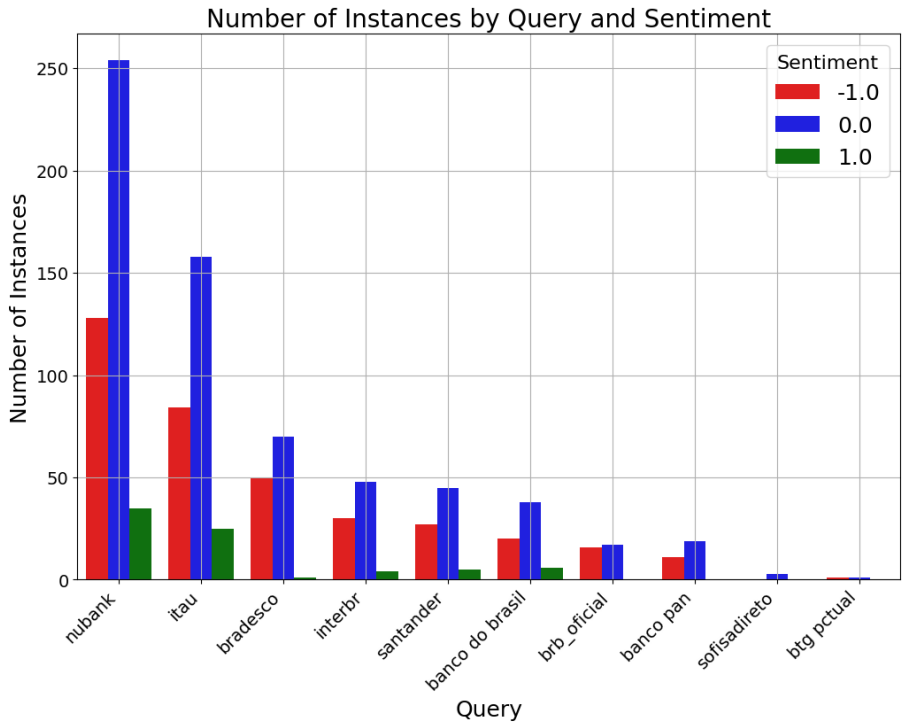


Figure 3. Relationship between the number of comments and the query used to search them

4.2. Sentiment Analysis Task

To evaluate the dataset's effectiveness for sentiment analysis tasks, it was used to train several models: BERTimbau, Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost. Due to the relatively small size of the labeled dataset, deep neural network approaches were not feasible. Therefore, we opted for classical machine learning methods such as XGBoost, SVM, and LR. Additionally, BERTimbau, a pretrained model for the Portuguese language, was also included to leverage its language-specific advantages.



(a) Number of Comments by Query and Sentiment

(b) Heatmap of the Cohen kappa be- tween the labels of each volunteer

Figure 4. Visualization on the data distribution and agreement between labels of the labeled data

The text data underwent preprocessing steps including noise removal, tokenization, and word embedding. The dataset was randomly split into training and testing sets, with 80% used for training and 20% for testing. For SVM, LR, and XGBoost, we used the

Bag of Words (BOW) for embedding, while BERTimbau utilized its embedding model. To evaluate the effectiveness of the models, we measured the metrics F1-score macro, accuracy, and Recall macro. The macro average was chosen to consider the effectiveness in all three classes. Table 2 presents the results indicating that BERTimbau achieved the best results, followed by SVM.

Table 2. Effectiveness of the Sentiment Analysis Models (BERTimbau, XGBoost, SVM e LR)

Model	F1-Macro	Accuracy	Recall-macro
BERTimbau	0,62	0,71	0,59
XGBoost	0,34	0,59	0,38
SVM	0,40	0,63	0,41
LR	0,33	0,59	0,38

5. Discussion

The primary objective of creating and labeling the dataset was to fine-tune Sentiment Analysis models and expand the corpus of labeled data in Portuguese. Our dataset, sourced from Twitter/X, consists of 375,912 comments, with 1,096 labeled for sentiment. We focused on comments about Brazilian banks due to their extensive use and popularity, ensuring thematic consistency and a sizable corpus. Unlike other datasets that rely on metadata like scores or "likes" or Large Language Models (LLMs) like ChatGPT, our dataset was labeled by five volunteers, providing a unique focus on the banking industry. Beyond these goals, the Brazilian Tweet Banks Dataset (B2T) provides valuable data for various business and market analysis applications.

To ensure data privacy, we have deliberately chosen not to share usernames, UserID and Texts by the Twitter/X data-sharing guidelines. To access the complete data, including the comments' text, users can use the provided Tweet IDs to retrieve the corresponding content via the Twitter/X API.

This dataset is particularly relevant to the banking sector and research related to training sentiment analysis models. Given the dataset's scope, it can be instrumental in discerning customer sentiments such as enthusiasm or frustration towards banking products or services. Potential applications and integration methods include Customer Feedback Analysis, Competitive Analysis, and Product and Service Improvement. These applications can leverage the dataset to gain insights into customer perceptions and market trends.

The results from the sentiment analysis models demonstrated that the dataset is effective for training such models. Among the tested models, BERTimbau achieved the highest accuracy, scoring 71%. These results could be enhanced by increasing the amount of labeled data. As more data are labeled, we anticipate that future research will show improved effectiveness across all models explored in our experiments.

While this study focused on using the dataset for fine-tuning Sentiment Analysis models, the Brazilian Tweet Banks Dataset (B2T) offers numerous other opportunities for experimentation and insights. For example, applying Topic Modeling techniques can

reveal the most discussed topics, products, or services, providing valuable market intelligence. Additionally, analyzing the presence of different banks on social media can offer insights into their marketing strategies. Understanding which banks have a stronger or weaker social media presence can inform competitive analysis and strategic planning, allowing businesses to refine their marketing approaches and enhance customer engagement.

While the dataset offers considerable potential for various analyses, it also presents limitations. One key issue is the class imbalance, where a more significant proportion of users express disapproval than satisfaction on social media, which could negatively impact the effectiveness of models trained on this data. Another challenge lies in the agreement among raters. Although there is moderate inter-rater agreement, this could be improved by refining the labeling methodology and providing more explicit instructions to raters.

Future work will address these limitations and explore additional aspects of the dataset, such as analyzing comment topics and applying advanced NLP techniques to gain deeper insights into how events and news influence the volume and sentiment of comments about banks. Moreover, incorporating machine translation for datasets in foreign languages will allow the training of new sentiment analysis models without significant loss of information. This approach will be integrated and compared with models trained without this step. Finally, enhancing sentiment labeling methods and expanding the dataset to cover a broader range of topics and sources will improve its utility and applicability for research and practical use.

6. Conclusion

Sentiment analysis has seen substantial growth and application across various domains, such as business popularity analysis, customer feedback analysis, and market research. The scarcity of labeled datasets has hindered the development of high-quality sentiment analysis models specifically for the Portuguese language. Our study created a new dataset curated for sentiment analysis in the Portuguese language. The dataset has proven helpful for training sentiment analysis models such as BERTimbau, XGBoost, Support Vector Machine, and Logistic Regression. The dataset can further be utilized for various analyses, such as topic modeling, to uncover additional insights and trends within the industry. Improvements can be made by increasing the number of labeled data, refining rater guidelines to enhance agreement scores, applying cross-validation to evaluate the models deeply, and addressing class imbalance through methods like SMOTE to oversample the minority class.

Acknowledgements

This project was supported by the brazilian Ministry of Science, Technology and Innovations, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Arquitetura Cognitiva (Phase 3), DOU 01245.003479/2024 -10.

References

Alves, M., Macedo, M., Ribeiro, J., Mancine, L., and Júnior, C. P. (2024). Sentimentos em cena: uma análise dos comentários em trailers de filmes da netflix brasil no youtube.

In *Anais do XIII Brazilian Workshop on Social Network Analysis and Mining*, pages 228–234, Porto Alegre, RS, Brasil. SBC.

Brum, H. B. and das Graças Volpe Nunes, M. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *CoRR*, abs/1712.08917.

Drus, Z. and Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

França, T., Gomes, J., and Oliveira, J. (2017). A twitter opinion mining gold standard for brazilian uprising in 2013. In *XXXII Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2017 Companion*, pages 182–192.

Hovy, E. and Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Liu, B. et al. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Mohanty, A. and Cherukuri, R. C. (2023). Sentiment analysis on banking feedback and news data using synonyms and antonyms. *International Journal of Advanced Computer Science & Applications*, 14(12).

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Plotnikov, A., Shcheludyakov, A., Cherdantsev, V., Bochkarev, A., and Zagoruiko, I. (2020). Data on post bank customer reviews from web. *Data in Brief*, 32:106152.

Saragih, M. H. and Girsang, A. S. (2017). Sentiment analysis of customer engagement on social media in transport online. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, pages 24–29.

Sousa, R. F. d., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Symposium in Information and Human Language Technology - STIL*. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.