Exploring Large Language Models for Portuguese Political Tweet Annotation[1]

Ricardo Cabral Penteado
Universidade de São Paulo
ricardo.penteado@usp.br

ABSTRACT

Investigating the annotation capabilities of LLMs, this research employs LlaMa2 (Bode7B, Bode13B) and text-davinci-003 to classify political tweets in Portuguese, based on a few-shot learning prompt aligned with at the "Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020" research codebook, which served as the ground truth for this investigation. Our analysis reveals varying degrees of alignment between LLM predictions and human annotations, underscoring the need for balanced datasets to improve model performance. Additionally, we identify the complexity involved in classifying neutral classes, characterized by highly nuanced and intricate discourse. To assess the model's difficulty in making accurate assessments, we employ confidence coefficients that measure the reliability of model predictions, offering insights into the consistency of human-generated annotations and measures of inter-annotator agreement, which quantify the level of agreement among human annotators, thus assessing the reliability and consistency of human-generated annotations. This study contributes to optimizing LLMs for processing politically charged language in Portuguese.

1 INTRODUCTION

In recent years, the emergence and advancement of Large Language Models (LLMs) have opened new avenues in the field of natural language processing, particularly in understanding and interpreting social media content. This paper explores the efficacy of LLMs in annotating Portuguese-language political tweets, an area that presents challenges due to the nuanced nature of political discourse and the linguistic complexity of the Portuguese language.

[1] Presented to FLS-6513 Processamento de Língua Natural Aplicada para Ciência Política e Análise de Políticas Públicas (2023), Universidade de São Paulo.

The motivation for this study stems from the growing need for efficient and accurate analysis of discourse on social media. Traditional methods of tweet annotation require substantial human effort and are often constrained by the availability of resources, both in terms of time and expertise. LLMs offer a promising alternative, with their ability to process vast amounts of data quickly and with reduced computational requirements compared to conventional models. However, their effectiveness in accurately capturing the subtleties of political discourse in Portuguese tweets remains largely unexplored.

This research aims to evaluate the performance of different LLMs in the unsupervised annotation of Portuguese political tweets. The models chosen for this study are "Bode7B", “Bode13B" based on LlaMa2 (Meta), and “text-davinci-003” (Openai), selected based on their varying capacities and architectures. The first experiment involves assessing inter-annotator agreement, where each model annotates the same set of tweets, with "Bode7B" and "Bode13B" providing labels and confidence coefficients, and “text-davinci-003” evaluating tweets with discrepant annotations.

The second experiment focuses on balanced classes, with an equal number of tweets assigned to each category, annotated by different types of models. This approach aims to mitigate class imbalance, a common issue in tweet datasets, and evaluate how well each model performs across a range of topics. For this experiment we used “Bode13B”, “text-davinci-003“ and "gpt-4"

To guide the annotation process, prompts based on the research project Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020's codebook were developed. These prompts are designed to align the models' annotations with specific research objectives, ensuring relevance and consistency.

The core of this study lies in comparing the annotations generated by the LLMs with existing manual annotations. This comparison will not only quantify the discrepancies between manual and automated annotations but also provide insights into the strengths and weaknesses of each approach.

Finally, the study will discuss the potential applications and implications of automated annotations in political analysis. It will explore how LLMs can be leveraged in political studies, the advantages they offer over manual annotations, and the challenges they pose, thereby contributing to the broader discourse on the role of artificial intelligence in social science research.

2 LLM EVALUATION

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) are characterized by their extensive parameter sizes, usually in the billions. These models undergo initial training with vast natural language datasets, notable examples being GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), and BLOOM (Scao et al., 2022). Their capability to perform remarkably on new tasks with only instructions, a process known as zero-shot in-context learning, is a significant feature. Post-pre-training enhancements, such as with T0 (Sanh et al., 2022) and FLAN (Wei et al., 2022), have been implemented to refine this zero-shot in-context learning. These models are fine-tuned across various tasks, exhibiting superior zero-shot capabilities compared to the original GPT-3 model. InstructGPT (Ouyang et al., 2022), an evolution of GPT-3, benefits from reinforcement learning based on human feedback (RLHF), showing improved adherence to instructions. A further development is ChatGPT (OpenAI, 2022), an iteration of InstructGPT fine-tuned with conversational datasets via RLHF. This enables ChatGPT to engage interactively with users, providing detailed answers and explanations to queries. The potential of LLMs to replace human evaluators and assist NLP researchers in text quality assessment is an area of exploration, given their proficiency in following instructions and generating feedback.

2.2 Annotation with LLMs

In recent research, the potential of Large Language Models (LLMs) like GPT-3.5 for data annotation is explored in "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators” by He et al. (2023) The paper introduces an innovative 'explain-then-annotate' approach, where LLMs are prompted to justify their choice of ground truth labels. This method has proven to enhance the annotation capabilities of GPT-3.5, yielding results that either surpass or are on par with human annotators in various tasks, demonstrating the potential of LLMs as viable substitutes for traditional crowdsourced annotators in specific contexts.

Similarly, "LLMAAA: Making Large Language Models as Active Annotators" by Ruoyu et al (2023) presents an advanced approach to improve LLMs' annotation efficiency in NLP tasks. The LLMAAA framework, integrating active learning, leverages prompt engineering and automatic reweighting to optimize both annotation and training. The framework's effectiveness is underscored through its success in tasks such as named entity recognition and relation extraction, where models trained with LLM-generated labels outperformed their teacher models.

"Can Large Language Models Be an Alternative to Human Evaluation?" by Chiang et al. (2023) delves into using LLMs, including GPT-3 and T0, for text quality assessment. This study shows that LLM evaluations align with expert human evaluations, suggesting LLMs as a cost-effective and efficient alternative in certain NLP tasks.

Additionally, "Is GPT-3 a Good Data Annotator?" by Ding et al., 2023, challenges conventional annotation methods by evaluating GPT-3's annotation performance in various NLP tasks. The research highlights GPT-3's capability in producing coherent, human-like text, positioning it as a promising alternative to traditional data annotation methods.

"From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks?" by Thapa et al., 2023, examines the feasibility of replacing human annotators with LLMs like ChatGPT in NLP tasks. The paper discusses the advantages and challenges of this approach, concluding that while LLMs offer time and cost efficiency, they may not entirely replace human annotators in all NLP scenarios.

This research employs LlaMa2 models (Bode7B, Bode13B) and text-davinci-003 to classify political tweets in Brazilian Portuguese. Our analysis demonstrates varying degrees of alignment between the model-generated annotations and human annotations, shedding light on the inherent challenges in processing politically charged language in this context.

2.3 Models

Llama2 is a collection of foundation language models ranging from 7B to 70B parameters, including checkpoints fine-tuned for chat applications. These models are part of the Llama 2-Chat series, optimized for dialogue use cases. They have shown to outperform other open-source chat models in most benchmarks tested, offering a potential substitute for closed-source models. The Llama2 model's development and refinement emphasize dialogue optimization and safety improvements, contributing significantly to the responsible development of large language models (LLMs).

Bode is a Portuguese language Large Language Model (LLM) developed from the Llama 2 model through fine-tuning on the Alpaca dataset by Garcia et al. 2023, which was translated into Portuguese by the creators of the Cabrita model. Designed for natural language processing tasks in Portuguese, such as text generation, automatic translation, and text summarization, Bode addresses the shortage of LLMs tailored for the Portuguese language. While models like LLaMa can respond to prompts in Portuguese, they often produce grammatical errors or generate responses in English. There are few Portuguese models available for free use, and to our knowledge, Bode is one of the rare models with 13 billion parameters or more specifically trained on Portuguese data, marking a significant advancement in language processing capabilities for Portuguese.

The text-davinci-003 model, released by OpenAI, represents a advancement in large language models. It is built on the InstructGPT framework and employs reinforcement learning with human feedback (RLHF) to better align with human instructions. This approach contrasts with its predecessor, davinci-002, which relied on supervised fine-tuning. The model is optimized to produce text that would be highly rated by humans, making it adept at producing higher quality writing, handling complex instructions, and generating longer form content.

2.4 LLM Evaluation

The research project "Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020” focuses on the political aspects of discourse surrounding COVID-19 vaccine sentiment on Twitter by Brazilian political elites in 2020. It aims to meticulously document the participation of these elites in debates about COVID-19 vaccines and vaccination. The project includes a detailed political elites dataset, categorizing candidates based on party affiliation, ideological positioning, and their alignment with or opposition to President Jair Bolsonaro's government. Specifically, it explores the extent to which candidates endorsed or aligned with Bolsonaro's ideologies and policies during the election campaign. This analysis of tweets is structured to assess the capabilities of large language models (LLMs) in annotating and understanding the nuanced political discourse reflected in these social media interactions.

The dataset of the referred project serves as a ground truth for evaluating the performance of large language models (LLMs) in political discourse analysis. This ground truth, established through expert analysis, provides a reliable benchmark against which the capabilities of LLMs can be assessed. By comparing the LLMs' annotations with these expert-generated annotations, we can gauge the accuracy and depth of understanding that LLMs bring to the complex task of interpreting political sentiment and alignment in social media discourse.

In this research, we utilized only 10% of the annotated dataset, ensuring that the balance between classes is maintained. This decision was driven by practical constraints, particularly the limited processing capacity and the substantial RAM requirements needed to run Large Language Models (LLMs). Additionally, budget constraints posed limitations on our capacity to leverage OpenAI's API for an exhaustive analysis of political tweet classifications.

2.5 First experiment

In this study, we adapted the concept of Inter-annotator Agreement (IAA) for evaluating the annotation capabilities of three advanced Large Language Models (LLMs): Bode7B, and Bode13B, with an inclusion of OpenAI's text-davinci-003 as a third 'virtual annotator.' This adaptation was designed to assess the consistency and reliability of these models in interpreting and annotating complex datasets. Bode models were evaluated against text-davinci-003.

Inter-annotator agreement (IAA) is a measure used to assess the reliability of an annotation process, crucial for ensuring the correctness of resulting annotations. It examines the extent to which different annotators provide consistent annotations for the same dataset. The premise is that, in the absence of a reference corpus for validating annotations, the reliability of the annotation process becomes a key factor. This involves evaluating how closely the annotations adhere to the guidelines and whether they represent a correct interpretation of the source material as intended by the experimenters. However, IAA assessment can be complex. The study bypassed the Kappa statistic, preferring direct assessment along dimensions like focus and polarity, due to Kappa's limitations across different datasets. Notably, annotator performance varied significantly, with some achieving high agreement levels. This variation underscores the need for thorough training to ensure consistent and reliable annotations.

In our methodology, the initial testing was conducted using the first two models, Bode (both 7B and 13B versions), to annotate the dataset with confidence coefficient. Following this, the annotations provided by these models were evaluated using OpenAI's text-davinci-003 as the third model, effectively functioning as a 'virtual annotator.' This approach was aligned with the adapted Inter-annotator Agreement (IAA) framework, allowing for a thorough assessment of the consistency and reliability of annotations across different Large Language Models (LLMs).

The utilization of confidence coefficients from the initial models, such as Bode (7B and 13B versions), served a crucial purpose in our methodology. Firstly, it allowed us to gauge the certainty levels of the model-generated annotations, enabling us to identify areas where the model exhibited higher or lower confidence. This information was invaluable in pinpointing potential trouble spots in the dataset that might require further human intervention or validation. Moreover, by incorporating confidence coefficients, we aimed to create a more transparent and interpretable annotation process, providing insights into the model's self-assessment of its performance. This approach facilitated a structured and systematic evaluation of the model's reliability and alignment with human annotators.

2.6 Second experiment

In the second experiment, we ensured that the dataset remained balanced, with an equal number of examples for each class. This approach was crucial for maintaining the integrity and representativeness of the dataset across different categories. For the analysis, we utilized GPT-4, text-davinci003 and Bode13B. We compared the results with the ground truth.

2.7 Prompt

In this research, we developed prompts based on the detailed classifications outlined in the codebook of "Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020” project, tailoring them to evaluate the annotation abilities of LLMs. The prompts encompass a sample of arbitrary text, a list of categories, and training data with examples of text samples and their assigned categories. These categories, such as 'Favorável', 'Desfavorável', and 'Neutro', are defined with specific keywords related to COVID-19 vaccines and vaccination sentiments.

2.8 Assessment metrics

In this research, we will employ the classification report as a key metric for evaluating the performance of our models. The classification report provides a comprehensive overview of the precision, recall, and F1-score for each class in the dataset, offering insights into the accuracy and reliability of the model's predictions. This report is particularly valuable in understanding how well the model distinguishes between different classes and its effectiveness in correctly classifying instances.

3 RESULTS AND DISCUSSION

3.1 First Experiment

In the first experiment, we selected 643 tweets for annotation, of which 522 received responses from the first two models. Among these, only 85 tweets exhibited annotation discrepancies between the models, indicating a substantial level of agreement in most cases. The Cohen's Kappa score for this experiment was calculated to be 0.4794339229214074. This moderate Kappa value suggests a fair but not strong agreement between the models, highlighting areas where their interpretation of tweets diverged, and underscoring the complexity of accurately annotating social media content, especially in the context of nuanced or ambiguous tweets.

In our research study, we conducted an analysis comparing the results of human annotators with those of Large Language Models (LLMs) in annotating the "Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020" dataset.

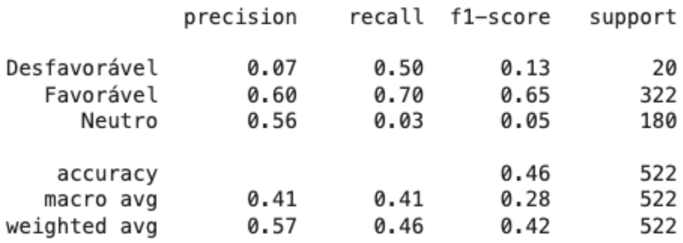
In the first aspect of this comparison, the human annotation was revisited during the course 'FLS6513 - Processamento de Língua Natural Aplicada para Ciência Política e Análise de Políticas Públicas (2023)', with 31 different annotators re-evaluating the dataset. Each tweet within the dataset was meticulously reviewed by two different annotators. This re-annotation process revealed a Cohen's Kappa coefficient of approximately 0.258 for this selected tweets. This value reflects a weak to moderate agreement among human annotators, suggesting variability and subjectivity in their interpretations of the tweets. The outcome highlights the challenges in ensuring consistency in human annotations, particularly in contexts where the content is nuanced or open to varied interpretations.

Conversely, the LLMs in our first experiment demonstrated a higher degree of agreement, with a Cohen's Kappa score of 0.4794339229214074. This score, while still in the moderate range, was notably higher than that of the human annotators. The LLMs' performance suggests a more consistent approach to annotation, possibly due to their programming and the extensive training data they have been exposed to. However, the score also indicates room for improvement, especially in terms of understanding and accurately interpreting the complex and often subjective nature of social media content.

The comparison between human annotators and LLMs in this context is illuminating. It not only reveals the current capabilities and limitations of LLMs in processing and understanding human language but also underscores the inherent complexities in annotating and interpreting social media content. As LLMs continue to evolve, these findings can guide further refinements in model development, aiming to bridge the gap between human and machine annotation capabilities, especially in complex and subjective domains like political discourse analysis.

Another pivotal evaluation criterion is the confidence coefficient. When comparing the outcomes, the "Bode7B" model demonstrates a notable level of confidence in categorizing tweets as "Desfavoráveis" and "Favoráveis," boasting mean confidence scores of approximately 0.818 and 0.847, respectively. Meanwhile, the "Bode13B" model extends its classifications to include a "Neutro" category but displays higher variability and uncertainty in this category, as evident from the lower mean confidence score of approximately 0.400. These findings suggest that while the models demonstrate notable confidence in classifying politically oriented tweets as favorable or unfavorable, the introduction of a neutral category introduces greater ambiguity, reflecting the challenges associated with assessing neutrality in political discourse.

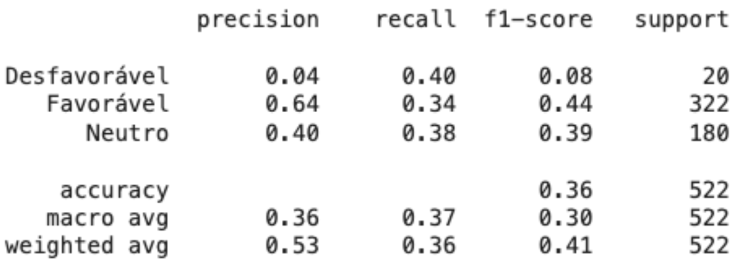
After evaluating the third LLM, we compared the results using the classification report provided. The model demonstrated the highest precision (0.60) and recall (0.70) for the 'Favorável' category, resulting in the highest F1-score (0.65) among the categories. In contrast, the 'Desfavorável' category had notably low precision (0.07), but a moderate recall (0.50), indicating a tendency of the model to misclassify other categories as 'Desfavorável.' The 'Neutro' category showed reasonable precision (0.56), yet very low recall (0.03), suggesting that the model struggled to identify neutral tweets accurately. Overall accuracy stood at 0.46, with a macro average F1-score of 0.28, pointing to an imbalance in the model's performance across different classes. The weighted average scores were somewhat higher due to the imbalance in class distribution, reflecting the model's better performance on the more represented 'Favorável' category.



```
              precision    recall  f1-score   support

Desfavorável       0.07      0.50      0.13        20
   Favorável       0.60      0.70      0.65       322
      Neutro       0.56      0.03      0.05       180

    accuracy                           0.46       522
   macro avg       0.41      0.41      0.28       522
weighted avg       0.57      0.46      0.42       522
```

Table 1: First Experiment vs. Ground Truth

When contrasting the ground truth with annotations solely from text-davinci-003 for all 522 tweets, a distinct performance profile becomes evident. For the 'Desfavorável' category, precision is very low at 0.04, suggesting that the model frequently mislabels other categories as 'Desfavorável.' However, its recall is at 0.40, indicating that it is somewhat capable of identifying 'Desfavorável' tweets when they occur. The 'Favorável' category shows better precision at 0.64 but lower recall at 0.34, meaning the model accurately identifies 'Favorável' tweets but misses many of them. 'Neutro' has a balanced precision and recall (0.40 and 0.38, respectively), with a moderate F1-score of 0.39. The overall accuracy of the model is 0.36, with a macro average F1-score of 0.30, and a weighted average F1-score of 0.41, reflecting the model's challenges in classifying tweets consistently with the ground truth.

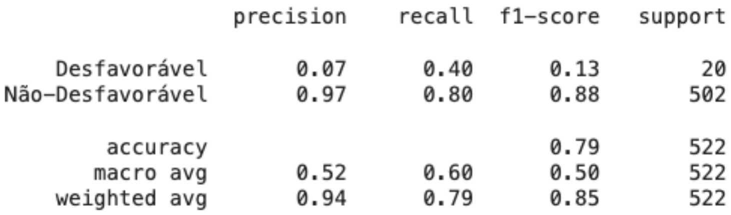


```
               precision    recall  f1-score   support

Desfavorável       0.04      0.40      0.08        20
   Favorável       0.64      0.34      0.44       322
      Neutro       0.40      0.38      0.39       180

    accuracy                           0.36       522
   macro avg       0.36      0.37      0.30       522
weighted avg       0.53      0.36      0.41       522
```

Table 2: text-davinci-003 vs. Ground Truth

The challenge posed by previous models in identifying and categorizing tweets in the neutral class is quite evident, particularly considering the models utilized. However, a significant improvement is observed when the text-davinci-003 model is employed on the entire corpus, as this category emerges more prominently only with the use of this specific model. Notably, the "Bode7B" model did not exhibit any neutral class. Recognizing this, we conducted an additional test that excludes the neutral class, focusing solely on the "unfavorable" and "non-unfavorable" categories.

If we consider the data in a binary fashion, categorizing tweets as either 'Desfavorável' or 'Não-Desfavorável,' the model exhibits a precision of 0.07 for 'Desfavorável,' suggesting it seldom correctly identifies tweets as such. However, it has a higher recall of 0.40, indicating it captures a fair portion of the 'Desfavorável' tweets but with a high rate of false positives. Conversely, the model performs significantly better for the 'Não-Desfavorável' category, with a high precision of 0.97 and recall of 0.80, leading to an impressive F1-score of 0.88. This indicates a strong ability to correctly identify and classify non-'Desfavorável' tweets. The overall accuracy is 0.79, reflecting a high level of correct classifications across the dataset when viewed in binary terms.



```
                   precision    recall  f1-score   support

    Desfavorável       0.07      0.40      0.13        20
Não-Desfavorável       0.97      0.80      0.88       502

        accuracy                           0.79       522
       macro avg       0.52      0.60      0.50       522
    weighted avg       0.94      0.79      0.85       522
```

Table 3: Binary Classification

3.2 Second Experiment

After assessing the performance of three distinct models on a dataset comprising 756 tweets, each class having an equal number of examples, we meticulously evaluated the provided performance metrics in comparison to a meticulously annotated ground truth dataset, meticulously curated by domain specialists. The models used in this study were employed to annotate the tweets through the process of few-shot learning. This approach involved utilizing the same prompts as those used in the previous experiment, facilitating a direct comparison of the results. These prompts served as guidelines for the models, directing them in classifying the messages according to the established categories.

Model	Favorável F1-Score	Desfavorável F1-Score	Neutro F1-Score	Micro Avg F1-Score	Macro Avg F1-Score
Bode13B	0.53	0.67	0.65	0.62	0.61
text-davinci-003	0.06	0.11	0.51	0.35	0.23
GPT-4	0.51	0.69	0.33	0.53	0.51

Table 4: Second Experiment

For the "Bode13B" model, it demonstrated reasonably strong precision (0.69) and recall (0.43) for the "Favorável" class, resulting in an F1-Score of 0.53, indicating a good alignment with the expert-annotated ground truth. Similarly, it exhibited solid performance for the "Desfavorável" class with a precision of 0.64, recall of 0.70, and an F1-Score of 0.67, suggesting a strong agreement with the specialists' annotations. However, for the "neutro" class, it had a lower F1-Score of 0.65, indicating the need for improvement in capturing the nuanced ground truth within this category. The micro-average F1-Score (0.62) provided an overall assessment of the model's performance, taking into consideration the distribution of the ground truth data.

Conversely, the "text-davinci-003" model exhibited extreme disparities in performance. It displayed very low precision (0.17) and recall (0.04) for the "favorável" class, resulting in an F1-Score of 0.06, which significantly deviated from the expert-annotated ground truth. However, it demonstrated high precision (0.82) and recall (0.95) for the "desfavorável" class, resulting in an F1-Score of 0.51, indicating a strong alignment with the specialist annotations in this category. Remarkably, it excelled in predicting the "neutro" class with a precision of 0.35, recall of 0.51, and an F1-Score of 0.35, closely mirroring the ground truth. The micro-average F1-Score (0.23) and macro-average F1-Score (0.23) revealed disparities among classes in comparison to the specialist-annotated data.

Finally, for the “GPT-4" model, it demonstrated strong precision (0.74) and recall (0.39) for the "Favorável" class, resulting in an F1-Score of 0.51, aligning well with the expert-annotated ground truth. Similarly, it excelled in predicting the "Desfavorável" class with a precision of 0.80, recall of 0.69, and an F1-Score of 0.45, demonstrating a strong agreement with specialist annotations. However, it exhibited relatively weaker performance for the "Neutro" class with an F1-Score of 0.33, suggesting potential room for refinement in this category. The micro-average F1-Score (0.53) and macro-average F1-Score (0.51) offered insights into the model's overall performance, considering the distribution of the ground truth data.

Certainly, considering that all classes have an equal number of examples, it's evident that the models' performances vary significantly in capturing the nuances within each class. The "Bode13B" model demonstrates relatively balanced performance across all classes, suggesting its ability to generalize well and provide consistent predictions across different polarity categories. On the other hand, the "text-davinci-003" model exhibits discrepancies, excelling in some categories while severely underperforming in others. This indicates a potential limitation in its capacity to handle the diversity of language expressions within the dataset. Lastly, the “GPT-4" model performs strongly in capturing favorable and unfavorable sentiments but struggles with the neutral category. These results emphasize the importance of model tuning and optimization to ensure more consistent performance across all classes when faced with a balanced dataset, while also highlighting the need for further investigation into the models' limitations in handling certain linguistic nuances.

In the context of the research, the annotation performance of Bode13B, text-davinci-003 and GPT-4 reveals nuanced capabilities and limitations. Bode13B maintained a robust balance between precision and recall across categories, with its F1-Scores indicating reliable performance, particularly in recognizing neutral sentiment. Text-davinci-003's precision was notably high in identifying 'Desfavorável' tweets; however, its low recall suggests a conservative approach, potentially missing several relevant annotations. Contrastingly, GPT-4 displayed strong precision for 'Favorável' tweets and excellent recall for 'Desfavorável', yet struggled with 'Neutro' tweets, as evidenced by its lower F1-Score. The disparity in Micro, Macro, and Weighted F1-Scores among these models underscores the complexity of annotating social media content and the challenges in achieving high accuracy in sentiment analysis. These findings highlight the need for tailored model training and refinement to enhance performance in specific annotation tasks within the realm of political discourse.

3.3 Discussion

When we delve into the metrics and class balance in the two experiments, we see a nuanced picture of LLM performance. In the first experiment, the moderate Cohen's Kappa score indicated a decent level of agreement between the models, but not without discrepancies. These discrepancies could partly stem from class imbalance, as models often find it more challenging to accurately predict minority classes.

In the second experiment, when comparing against the ground truth, the LLMs showed variation in precision and recall, with some models better at identifying certain sentiments than others. The precision for 'Desfavorável' sentiments was significantly low, which might be due to the fewer instances of this class in the dataset, leading to less exposure during model training. The higher precision and recall for 'Favorável' sentiments in the Bode13B and GPT-4 models suggest a better grasp on this more prevalent category.

Overall, the class imbalance seems to have a pronounced effect on model performance. The LLMs were more adept at predicting classes with more examples, underscoring the importance of balanced datasets for training to achieve equitable sensitivity across categories. This aspect is critical for enhancing the models' utility in accurately reflecting diverse political discourse nuances.

5 CONCLUSION

The exploration of Large Language Models (LLMs) in annotating Portuguese political tweets has provided valuable insights into the capabilities and limitations of current AI technologies in understanding complex, nuanced human communication. Through the comparative analysis of Bode7B, Bode13B, and text-davinci-003 models, it became evident that while LLMs can process vast quantities of textual data, their performance is highly dependent on the dataset's balance and the representativeness of each class. Additionally, the challenge of classifying the "neutral" label became apparent, mirroring the discrepancies observed in the human re-annotation process conducted during the course.

The discussion of results from the experiment involving Inter-annotator Agreement (IAA) reflects on the intricacies of LLMs' performance in annotating political tweets. The moderate IAA score observed suggests a fair level of agreement among the models, but also indicates room for improvement. This outcome highlights the complexity inherent in automated sentiment analysis, especially in politically charged contexts where nuances and subtleties of language are critical. The discrepancies in annotations between the models could be attributed to differences in their training datasets, underlying algorithms, or their handling of nuanced language. The findings point towards the need for enhanced training methods and more representative datasets to improve the alignment of LLMs with human judgment in complex annotation tasks.

Incorporating Cohen's Kappa into the evaluation of our LLMs provides a nuanced understanding of model agreement, but it is crucial to recognize its limitations in reflecting accuracy against the ground truth. While Kappa offers insight into the consistency between different annotators or models, a moderate or low score might indicate a divergence from the ground truth. This discrepancy suggests a potential need for further model optimization or a reevaluation of the annotation guidelines. Therefore, while Kappa is an important metric in our analysis, it should be interpreted in conjunction with other performance measures to gain a comprehensive understanding of the models' accuracy and reliability in complex sentiment analysis tasks.

Our experiments have shown that models can achieve a moderate level of accuracy in identifying favorable sentiments, a task made easier by the preponderance of such sentiments in the training data. However, they struggle with less represented sentiments like the unfavorable ones, pointing to a critical need for diverse and balanced datasets for training. This balance is not merely a technical requirement but a prerequisite for models to provide fair and unbiased analyses, reflecting the true spectrum of political discourse.

Looking forward, this study opens doors to future research avenues, with a specific emphasis on enhancing the utility of Large Language Models (LLMs) in the annotation of political data. The exploration of broader sociopolitical applications of LLMs holds considerable promise, offering an exciting and fertile ground for upcoming scholarly investigations.

6 REFERENCES

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." Advances in neural information processing systems, 33 (2020): 1877–1901.

Chiang, Cheng-Han, and Hung-Yi Lee. "Can Large Language Models Be an Alternative to Human Evaluations?" 2023. https://doi.org/10.48550/ARXIV.2305.01937.

Ding, Bosheng, et al. "Is GPT-3 a Good Data Annotator?" 2022. https://doi.org/10.48550/ARXIV.2212.10450.

Garcia, Gabriel Lino et al. "BODE-7b." 2023. Hugging Face, https://huggingface.co/recogna-nlp/bode-7b-alpaca-pt-br. doi:10.57967/hf/1298.

He, Xingwei, et al. "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators." 2023. https://doi.org/10.48550/ARXIV.2303.16854.

Le Scao, Teven, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).

OpenAI. "Chatgpt: Optimizing language models for dialogue." Accessed on January 10, 2023.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res., 21(140) (2020): 1–67.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. "Multitask prompted training enables zero-shot task generalization." In International Conference on Learning Representations (2022).

Thapa, Surendrabikram, et al. "From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks." Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, vol. 2023, 2023, p. 15.

Wei, Jason, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. "Finetuned language models are zero-shot learners." In International Conference on Learning Representations (2022).

Zhang, Ruoyu, et al. "LLMaAA: Making Large Language Models as Active Annotators." 2023. https://doi.org/10.48550/ARXIV.2310.19596.

ANNEX I - PROMPT

~~~
FEW_SHOT_CLF_PROMPT_TEMPLATE = """
Você receberá as seguintes informações:

1. Uma amostra de texto arbitrário. A amostra é delimitada com três crases.
2. Lista de categorias às quais a amostra de texto pode ser atribuída. A lista é delimitada com colchetes. As categorias na lista são cercadas por aspas simples e separadas por vírgulas.
3. Exemplos de amostras de texto e suas categorias atribuídas. Os exemplos são delimitados com três crases. As categorias atribuídas estão incluídas em uma estrutura semelhante a uma lista. Esses exemplos devem ser usados como dados de treinamento.

Categorias de postura:
- 'Favorável': Classifique um tweet como "Favorável" se promover vacinas e vacinação contra a COVID-19. Palavras-chave: Desenvolvimento, Aprovação, Urgência, Vacinação, Confiança, Ciência, Ensaios Clínicos, Avanços, Acordos, Campanhas, Mandatos, Restrições.
- 'Desfavorável': Classifique um tweet como "Desfavorável" se expressar posições desfavoráveis em relação à vacinação e às vacinas contra a COVID-19. Palavras-chave: Criticam, Questionam, Aprovação, Aquisição, Adoção, Desencorajam, Marcas de Vacinas, Eficácia, Confiança, Efeitos Colaterais, Segurança, Base Científica, Organizações Internacionais de Saúde, Empresas Farmacêuticas, Laboratórios, Instituições de Saúde, Agências Reguladoras, Restrição de Atividades, Cobertura Vacinal, Isolamento Social, Propagação do Vírus, Vacinação Obrigatória, Passaportes de Vacinação.
- 'Neutro': Classifique um tweet como "Neutro" se referir à vacinação e/ou vacinas contra a COVID-19 sem expressar julgamentos de valor.

Realize as seguintes tarefas:

- Identifique a qual categoria o texto fornecido pertence com a maior probabilidade.
- Atribua o texto fornecido a essa categoria.
- Forneça sua resposta contendo uma única chave 'label' e um valor correspondente à categoria atribuída. Não forneça nenhuma informação adicional.


Lista de categorias: {labels}

Dados de treinamento:
```Recife vai ter uma política municipal de vacinação robusta, para a gente poder vacinar as pessoas e com isso garantir as retomadas de maneira permanente ou ao longo do tempo. @cbnrecife #Eleicoes2020```, 'Favorável'
~~~

```Não vai ter Bolsonaro boicotando a Ciência! Anvisa autoriza retomada de testes do CoronaVac. São as pequenas luzes no fim do túnel essas notícias. Esperança sempre!```, 'Favorável'

```Hoje é um dia histórico, uma grande vitória da ciência! Margaret Keenan, de 90 anos, se tornou a primeira pessoa no mundo a ser vacinada contra covid-19. Que hoje seja o dia em que a humanidade começou a vencer a guerra contra esse vírus terrível…```, 'Favorável'

```Anvisa interrompe testes da vacina chinesa após ‘evento adverso grave’ em voluntário brasileiro. A VACHINA dando seus problemas 😅.conexaopolitica.com.br/brasil/anvisa-...```, 'Desfavorável'

```A vacina mais badalada no Brasil, AstraZeneca + Oxford apresentou resultados controversos de eficácia: 90% para meia dose e 62% para uma dose! #rodrigopaivanovo```, 'Desfavorável'

```Não estou falando de velocidade nas estradas. A pauta é uma vacina que ainda não foi suficientemente testada.```, 'Desfavorável'

```Os desafios científicos e tecnológicos a nossa frente vão muito além da descoberta e produção em massa de uma vacina que funcione. Ainda temos muitas perguntas sem resposta, em relação a essa doença. E encontrar uma solução para isso é fundamental.```, 'Neutro'

```Como eu registrei em entrevista à TV Guará, o debate político firme das ideias não pode ser impedinditivo da solidariedade no momento da doença, sobretudo essa Covid que ainda não tem vacina. Não se tripudia com doença dos outros. Pleno restabelecimento ao @carlosmadeiraof```, 'Neutro'

```Também é muito importante a transparência na compra dessas vacinas. Como deputado estadual eu vou fiscalizar a ação do Governo Do Estado pra garantir saúde e respeito ao dinheiro público.```, 'Neutro'

Exemplo para classificação: ```{x}```

Sua resposta:
"""

Additional information and access to the codes can be found in https://github.com/Penteado89/LLMs_annotation