Predicting *Big Brother Brasil 2024* Evictions Through Sentiment Analysis of Tweets

Stanford CS224N Custom Project

Laura Fiuza Dubugras
Department of Computer Science
Stanford University
lfiuza@stanford.edu

Abstract

This project endeavors to forecast the next eviction outcome in Brazil's foremost reality TV show. With an estimated weekly viewership in the millions and a track record of catapulting contestants into cultural prominence, the show holds substantial influence over public discourse. Through the lens of sentiment analysis applied to Twitter data, we delve into the intricate interplay of language, culture, and technology, aiming to unveil how public sentiment shapes real-world outcomes. Our approach capitalizes on the robustness of BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art language model, which we fine-tune to scrutinize tweets pertaining to participants in the current edition of the show. Notably, our model accurately predicted the eviction outcome in 2 out of 3 analyzed instances, underscoring its potential utility in forecasting public opinion-driven events.

1 Key Information to include

- Mentor: Caleb Ziems
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

In Brazil, one of the largest reality TV shows commanded the attention of 100 million people in the first month of the current edition, representing approximately half of the Brazilian population Vaquer (2024). This viewership also has financial implications for the partners that sponsor the show, with one of the partners registering 20% increase in its e-commerce sales Vaquer (2024). *Big Brother Brasil* has consistently shaped conversations, influenced pop culture and catapulted contestants into the limelight, with previous participants going off to become famous singers, entrepreneurs and brand affiliates.

As the stakes rise and tensions mount within the competition, the question on everyone's mind becomes: who will be the next to be voted off by the public? In this pursuit, I turn to BERT, a state-of-the-art pre-trained model renowned for its ability to grasp contextual nuances and semantic meanings within text.

3 Related Work

Sentiment Analysis in Election Context Chandra and Saini (2021) proposed a framework for modeling the U.S. general elections using BERT, specifically the 2020 U.S. presidential elections.

Khan et al. (2023) similarly introduced ElecBERT, a model specifically designed for sentiment analysis in election-related tweets.

Sentiment Analysis in News and Social Media Kawintiranon and Singh (2022) introduced PoliBERTweet, and highlighted the importance of domain-specific language models for tasks such as political misinformation analysis and election public opinion analysis. Azzouza et al. (2019) proposed a framework for Twitter sentiment analysis based on pre-trained BERT representations, which is more general than those related to elections.

4 Approach

The first step in my approach is data pre-processing, where we curate and prepare the dataset for model training. Each tweet is accompanied by annotations indicating sentiment towards the contestants and their likelihood of eviction, providing valualble labeled data for training our predictive model.

During pre-processing, I also tokenize the text using the `BERT` tokenizer, encoding each tweet into numerical representations suitable for input into the neural network model. More pre-processing cleanup and annotation details are expanded in the **Data** section.

Central to my approach is the utilization the existing pre-trained language model developed by Google. BERT's architecture consists of multiple layers of bidirectional transformers, allowing it to capture rich contextual information and semantic meanings within text data.

5 Experiments

5.1 Data

One of the challenges I faced was that originally I wanted to train on all eviction outcomes of recent history, traversing many different editions in the last several years. However, the Twitter API has some restrictions that only allow you to fetch tweets from less than 7 days ago. That added a challenge to the collection of tweets.

One other consideration I had to overcome was that ideally we'd have the labels represent eviction outcomes instead of sentiment. However, if I had opted for that route, I would have ended up with a very unbalanced dataset. For instance, if I am only analyzing 3 weeks and there are 3 eviction outcomes, there are essentially only 3 labels. That would make the model not very good at generalizing for future eviction polls because it would probably output one of the 3 previous eviction outcomes it "knows", which is completely inappropriate because by definition if someone has been evicted from the program they will not be up for eviction in the future.

My approach to the classification problem involved the following steps. Collected 500 tweets every Monday for three weeks totaling 1,500 tweets. Mondays are unique to *Big Brother Brasil* because that is when an eviction poll is ongoing. Found tweets that spoke about the show using the official hashtag `bbb24`.

Tweets were cleaned to remove noise, which included removing unnecessary hashtags, links and parts of the text that indicated that it was a retweet (e.g. "RT:").

I tagged each of the tweets with an array to indicate which of the three participants that are up for eviction are being referenced in each of the tweets, if at all. In the script that detected mentioning of these participants by name, we also had to check for common nicknames people use to refer to these participants.. For instance, one of the participants is called "Lucas" but his other common nickname is "Buda" so we searched for that string in the tweets as well.

I recruited 8 annotators to annotate each of the tweets as negative, positive or neutral as it pertains to each of the participants, respectively -1, 1 and 0 scores. If a tweet mentioned more than one person, I "repeated" the tweet as many times as there are participants mentioned and then asked the annotators to label the tweet with respect to the given participant.

After collecting all of the annotations, I ran the Fleiss-Kappa method to evaluate whether the annotations were adequate to be used. The Fleiss-Kappa came back with a score of 0.3429, which represents a Fair agreement between annotations. Future work may include using GPT-4 in order to annotate the tweets and comparing results to this report's predictions.

Later on the tweets were compiled to compose the multi-label, multi-class dataset, where the inputs are the Tweets in text form and the output is an n-dimensional array where n is the number of participants on the show and each element of the array is one of 3 options: -1, 1, and 0. More details to follow in the **Experiments** section.

The elements of each array were picked using majority vote of each of the labels that the 8 annotators output. If there was a tie, the label 0 was used.

5.2 Evaluation method

After fine-tuning the model with 90% of the dataset, the model was evaluated against the remaining 10% reserved validation set. I collected the labels that were predicted and calculated the percent of tweets that were negative toward each given participant that is up for eviction. The highest rate of negativity toward a participant was considered as the "predicted" participant that would be chosen to be evicted.

5.3 Experimental details

I utilized the pre-trained BERT model for sequence classification. The BERT model was fine-tuned using the `BertForSequenceClassification` class from *Hugging Face* `transformers` library. The model was initialized with the `bert-base-uncased` pre-trained weights and configured to predict among 8 possible eviction labels.

The tweets were tokenized using the BERT tokenizer with maximum sequence length set to 128 tokens and special tokens `CLS` and `SEP` addd for classification.

The number of epochs was 3, batch size was 16 per device for training and 64 per device for evaluation. The learning rate was initialized at `5e-5` with 500 warmup steps and 0.01 weight decay.

5.4 Results

Please see the **Appendix** for the breakdown of all metrics related to the predictions. Tables 2, 3 and 4 show breakdown of quantitative results for the first week's eviction poll; tables 5, 6 and 7 do so for the second week and tables 8, 9 and 10 do so for the third week.

Tables 2, 5, and 8 show the breakdown of the loss progression as the model goes through training, its associated epoch and learning rate. All tables show a decrease in loss values, which shows a sign of some learning.

Tables 3, 6, and 9 show the breakdown of the evaluation statistics, including evaluation loss and overall accuracy.

Tables 4, 7, and 10 show the sum of the predicted labels that were deemed negative from the validation set of tweets. If we theorize that the highest sum represents the most likely evicted contestant due to high negative sentiment found in tweets, we have as predictions that Rodriguinho, Davi and Yasmin will be chosen to be evicted.

6 Analysis

The model predicted 2 out of 3 evictions correctly. It did not surprise me that it did not predict the second week correctly. Davi is a long time favorite in this edition and happened to be involved in a expulsion of a candidate the week he was up for eviction. The expulsion happened because of another candidate, Wanessa, who was involved in an aggression act against Davi and did not obey the program rules where you are not allowed to engage in any aggression acts towards other contestants while you are participating in the program. Wanessa's expulsion was very controversial and was the first time that Davi's permanence in the program was up for debate, since all other weeks he was considered the strongest and most loved contestant in the program by a very large margin.

The model's main fault is that it does not have this context of the trajectory of the contestant's likeability in the program over the course of the edition, so the snapshot of a single day's tweets leading up to a public election, which is the main source of data for this model, is flawed. Even though there was indeed a large number of negative sentiment tweets pertaining to Davi in the second week, since people claimed the aggression act he reported was not as big of deal as he made it out to be, there was still a longtime favoritism attributed to him that kept him competitive in the program and ultimately led to another participant, Michel, to be evicted.

Moreover, the evaluation accuracy of the first week's model was pretty good, at around 87%, which the other two week's models decreased in accuracy substantially, at 69& and 53%, respectively. This demonstrates a high variability in accuracy, and thus probably a lot of room for improvement in future work.

7 Conclusion

This project dove into the fascinating world where technology intersects with culture and entertainment. Our findings revealed promising results, especially with so few tweets collected to populate the dataset. However, it's crucial to acknowledge the limitations encountered during the project The model's lack of contextual understanding regarding contestant's long-term likeability within the show proved to be a significant challenge. While negative sentiments towards certain contestants were detected, the model failed to account for the broader narrative surrounding their popularity, leading to an incorrect prediction.

Moving forward, there are many areas of improvement: a) time-series analysis could provide deeper insights into the evolving dynamics of public sentiment throughout the show's duration. Additionally, expanding data collection efforts to include other editions, as well as other social media platforms, such as Instagram could further enrich the dataset and enhance the model's performance. There is also room to incorporate feature engineering techniques such as including likes pertaining to comment, tweet or retweet, or even capturing significant events within the show that often affect public sentiment, such as conflicts between contestants, challenges won, romantic relationships. These could offer more nuanced predictions.

References

Noureddine Azzouza, Karima Akli-Astouati, and Roliana Ibrahim. 2019. Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. In *Advances in Intelligent Systems and Computing*.

Rohitash Chandra and Ritij Saini. 2021. Biden vs trump: Modeling us general elections using bert language model. In *IEEE*.

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: A pre-trained language model for analyzing political content on twitter. In *ACL Anthology*.

Asif Khan, Huaping Zhang, Nada Boudjellal, Arshad Ahmad, and Maqbool Khan. 2023. Improving sentiment analysis in election-based conversations on twitter with elecbert language model. In *Computers, Materials Continua*.

Gabriel Vaquer. 2024. Bbb 24: Com um mês no ar, reality vira hit entre adolescentes e chega a 100 milhões de pessoas. In *Folha*.

A Appendix

Epoch	Loss	Grad. norm.	Learning rate
0.34	0.7022	4.5326	1e-06
0.69	0.6733	4.4920	2e-06
1.03	0.6133	3.9819	3e-06
1.38	0.5415	3.7568	4e-06
1.72	0.4423	3.4342	5e-06
2.07	0.3899	3.3258	6e-06
2.41	0.3017	2.2374	7e-06
2.76	0.2522	1.9831	8e-06

Table 1: Training loss progression Week 1 (Fernanda v. Rodriguinho v. Lucas)

Eval. loss	0.1656
Eval. overall accuracy	0.8667
Eval. runtime	0.228
Eval. samples per second	219.294
Eval. steps per second	4.386
Epoch	3.0

Table 2: Evaluation set loss statistics Week 1 (Fernanda v. Rodriguinho v. Lucas)

Contestant	Sum of negative predicted labels
Fernanda	0
Rodriguinho	18
Lucas	1

Table 3: Predicted evicted contestant based on sum of negative predicted labels: Rodriguinho

Epoch	Loss	Grad. norm.	Learning rate
0.34	0.7639	4.5741	1e-06
0.69	0.7275	4.3488	2e-06
1.03	0.6430	4.0642	3e-06
1.38	0.5632	4.3328	4e-06
1.72	0.5109	2.0724	5e-06
2.07	0.4575	3.1367	6e-06
2.41	0.4140	3.1827	7e-06
2.76	0.3590	2.6619	8e-06

Table 4: Training loss progression Week 2 (Davi v. Alane v. Michel)

Eval. loss	0.2674
Eval. overall accuracy	0.6867
Eval. runtime	0.2317
Eval. samples per second	215.755
Eval. steps per second	4.315
Epoch	3.0

Table 5: Evaluation set loss statistics Week 2 (Davi v. Alane v. Michel)

Contestant	Sum of negative predicted labels
Davi	50
Alane	1
Michel	6

Table 6: Predicted evicted contestant based on sum of negative predicted labels: Davi

Epoch	Loss	Grad. norm.	Learning rate
0.34	0.7718	4.8447	1e-06
0.69	0.7481	4.8299	2e-06
1.03	0.6536	4.4405	3e-06
1.38	0.6027	4.5384	4e-06
1.72	0.5552	4.9571	5e-06
2.07	0.4913	3.4345	6e-06
2.41	0.4127	3.3880	7e-06
2.76	0.3267	2.6162	8e-06

Table 7: Training loss progression Week 3 (Lucas v. Yasmine v. Isabelle)

Eval. loss	0.2894
Eval. overall accuracy	0.5267
Eval. runtime	0.2342
Eval. samples per second	213.533
Eval. steps per second	4.271
Epoch	3.0

Table 8: Evaluation set loss statistics Week 3 (Lucas v. Yasmin v. Isabelle)

Contestant	Sum of negative predicted labels
Lucas	29
Yasmin	50
Isabelle	3

Table 9: Predicted evicted contestant based on sum of negative predicted labels: Yasmin