Analyzing the Emotional Tone of Donald Trump's Tweets Before and After the 2016 U.S. Presidential Election: A Neural Sentiment Classification Approach

Antonino Pio Lupo

AI Usage Disclaimer

Parts of this project have been developed with the assistance of OpenAI's ChatGPT (GPT-4). This model was employed for technical development purposes, including portions of the code and LaTeX formatting, as well as for enhancing the clarity, structure, and language of the written content.

All content produced with AI assistance has been carefully reviewed, edited, and validated by me. The creative process, critical decision-making, and all analytical work were entirely conducted by me. Generative AI tools were used strictly as support instruments to aid development and surface potential insights. I take full responsibility for the final content and its accuracy, relevance, and academic integrity.

Introduction

Social media platforms like Twitter have transformed the way political figures engage with the public, allowing for direct, real-time interaction with a global audience. Among these figures, former U.S. President Donald Trump stands out for his prolific and often controversial use of Twitter as a primary means of communication.

Tweets offer a valuable source of data due to their accessibility, brevity, and immediate nature. For Trump, they served not only as a tool for broadcasting policies and opinions but also as a window into his emotional and rhetorical stance over time. This makes his tweets a compelling subject for sentiment analysis.

In this project, we aim to train neural networks models to automatically classify the sentiment expressed in Trump's tweets, identifying the emotional tone of each of them. Beyond simple classification, the broader goal is to analyze sentiment trends over time and investigate whether there has been a noticeable shift in the tone of Trump's communication during his presidency.

Research Question and Methodology

This project aims to answer the following research question:

Is there a significant difference in the emotional tone of Donald Trump's tweets before and after becoming president?

To investigate this, we define the key temporal boundary as November 8, 2016, the date of Trump's first election victory. The dataset of tweets spans from 2009 to 2021, ending with the suspension of Trump's Twitter account.

Modeling Pipeline

To analyze this question, we design a multi-stage emotion classification pipeline that enables both high-level sentiment differentiation and fine-grained emotional analysis. The classifier is trained on the GoEmotions dataset, a widely-used emotion classification corpus compiled by Google. It contains 58,000 English Reddit comments, each annotated with one of 27 emotion categories or a neutral label. The dataset offers high annotation quality and emotional diversity, making it suitable for transfer learning to political text domains.

Specifically, we adopt a three-model training strategy:

Sentiment-level Classification

In the first stage, we train a model to classify tweets into positive, negative, or neutral sentiments. This task is conducted in two phases: an initial model is trained on the full dataset, followed by a second iteration with undersampling to address class imbalance.

Full Emotion Classification

The second model is trained on all 28 emotion categories defined in the GoEmotions dataset.

Fine-Grained Emotion Classification (Excluding Neutral)

The final model is trained using the same categories, but excluding the neutral class.

Category	Emotions
Positive	admiration, amusement, approval, caring, curiosity, desire, excitement, gratitude, joy, love, optimism, pride, relief
Negative	anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
Neutral	confusion, realization, surprise, neutral

Table 1: Categorization of emotions used for fine-grained emotion classification.

To standardize and sanitize the input text before tokenization, a dedicated preprocessing function was implemented. This procedure involves:

- Removal of non-alphanumeric and non-standard characters;
- Expansion of contractions (e.g., "don't" $\rightarrow$ "do not");
- Replacement of URLs with the token `[url]`;
- Elimination of numeric-only content and redundant whitespace;
- Conversion to lowercase for case insensitivity.

After preprocessing, texts were tokenized using a RoBERTa tokenizer to prepare them for model input.

Each classification model was trained using a RoBERTa-based transformer architecture, fine-tuned for the respective classification tasks. The dataset was split into training (81%), validation (9%), and test (10%) subsets, with the test set reserved for final evaluation. Performance metrics include accuracy, precision, recall, and F1-score, allowing for a comprehensive evaluation of model behavior across imbalanced and multi-class settings.

Performance Comparison

Models 1 and 2

Model 1, trained on three sentiment classes (positive, negative, neutral), demonstrates a moderate accuracy of 0.64. In terms of precision, recall, and F1-score, it performs reasonably well but exhibits an imbalanced focus, with a higher precision and recall for the negative class (0.72 and 0.71, respectively) compared to the neutral (0.57 precision, 0.63 recall) and positive classes (0.58 precision, 0.55 recall). This could indicate that the model tends to be more successful in predicting the negative sentiment but struggles with the more subtle neutral and positive sentiments.

Model 2, which utilizes undersampling to balance class distribution, achieves a slightly higher accuracy of 0.66. This improvement is reflected across the classes, particularly in the negative class (0.78 precision, 0.69 recall) and neutral class (0.54 precision, 0.77

recall), though the positive class still lags (0.62 precision, 0.54 recall). Despite a slight drop in the performance of the positive class, the overall better balance between precision and recall, especially for the neutral and negative classes, makes Model 2 more robust.

Recommendation: Model 2 is superior overall, particularly because its undersampling strategy allows for more balanced class predictions, which is critical for real-world applications where the distribution of sentiments often skews. The marginal improvements in accuracy, precision, recall, and F1-score suggest it would be the better choice for deployment.

Models 3 and 4

Model 3, trained on the GoEmotions dataset with 27 emotion categories plus neutral, shows a very low accuracy of 0.43. Despite this, the model performs relatively well for some emotions, such as love (0.61 F1-score) and amusement (0.65 F1-score), but struggles with others. This suggests the model's difficulty in distinguishing between more complex or nuanced emotions, a common challenge when working with a large number of categories. These results underperform the current state-of-the-art benchmarks reported in [1], where the authors achieved significantly higher scores using a fine-tuned BERT model. Notably, their approach excludes the neutral class from the outset in the fine-grained setup, diverging from the strategy adopted here.

Model 4, also trained on the same 27 emotion categories but excluding neutral, performs similarly with an accuracy of 0.44. While its F1-scores across emotions are slightly higher (notably gratitude at 0.71), it faces similar difficulties with rare or subtle emotions. The absence of a neutral class may contribute to the model's focus on more emotionally charged sentiments, possibly leading to biases in predictions.

Discussion: Both models are highly granular, targeting specific emotions, but their low accuracy indicates that such fine-grained categorization comes at the cost of performance. For sentiment analysis tasks, where broader classes like positive, negative, and neutral are more practical, these models may not be the most efficient choice. However, they could be beneficial in applications where understanding subtle emotional states is paramount, such as personalized sentiment analysis in social media analysis.

The RoBERTa model employed in this study for fine-grained emotion recognition also underperforms when compared to the bidirectional LSTM baseline reported in [1]. However, RoBERTa exhibits stronger performance in coarse-grained sentiment classification tasks. A key methodological distinction must be noted: this study adopts a three-class sentiment framework (positive, neutral, negative), whereas the [1] benchmark includes a fourth "ambiguous" category for emotions that defy clear positive or negative labeling and do not fit within the neutral category. This difference may partially account for the observed performance discrepancies.

Table 2: Emotion Classifiers Performances

Classification Level	Model	Accuracy	Precision (weighted-avg)	Recall (weighted-avg)	F1-Score (weighted-avg)
Coarser Classification	Model 1	0.64	0.64	0.64	0.64
	Model 2	0.66	0.67	0.66	0.66
Fine-Grained Classification	Model 3	0.43	0.41	0.43	0.41
	Model 4	0.44	0.43	0.44	0.43

Temporal and Sentiment Trends in Trump's Tweets

Dataset and Exploratory Data Analysis

The dataset used for the analysis comprises 56,571 tweets authored by Donald Trump between 2009 and 2021:

- `id`: Unique numeric identifier assigned to each tweet.
- `text`: Full textual content of the tweet.
- `is_retweet`: Boolean indicating whether the tweet is a retweet (TRUE or FALSE).
- `is_deleted`: Boolean indicating whether the tweet was deleted.
- `device`: String specifying the device or platform used to post the tweet.
- `favorites`: Integer count of likes the tweet received.
- `retweets`: Integer count of retweets the tweet received.
- `datetime`: String representation of the exact posting date and time.
- `is_flagged`: Boolean indicating if the tweet was flagged.
- `date`: Date-only timestamp in datetime format (YYYY-MM-DD).

During preprocessing, tweets containing the token `[url]` were removed from the dataset. These entries were considered sentimentally uninformative, as the token represents a hyperlink whose content is inaccessible and thus provides no usable context for sentiment analysis. This step reduced the dataset from 58,011 to 40,694 entries, contributing to a cleaner and more meaningful corpus.

Before delving into the core sentiment analysis, I conducted an exploratory data analysis to gain a deeper understanding of the dataset and the broader context of Trump's tweets.

As a starting point, a temporal analysis reveals that Trump's Twitter activity begins to intensify around 2013, reaching a significant peak during the lead-up to the 2016 U.S. presidential election. Following his electoral victory, tweet volume declines but remains relatively stable until the onset of the COVID-19 pandemic and the subsequent 2020 election campaign. Interestingly, despite this decrease in posting frequency after taking office, Trump's online presence becomes markedly more viral—suggesting that

the influence of his tweets amplified significantly during his presidency, likely due to the increased visibility and institutional weight of the presidential platform.

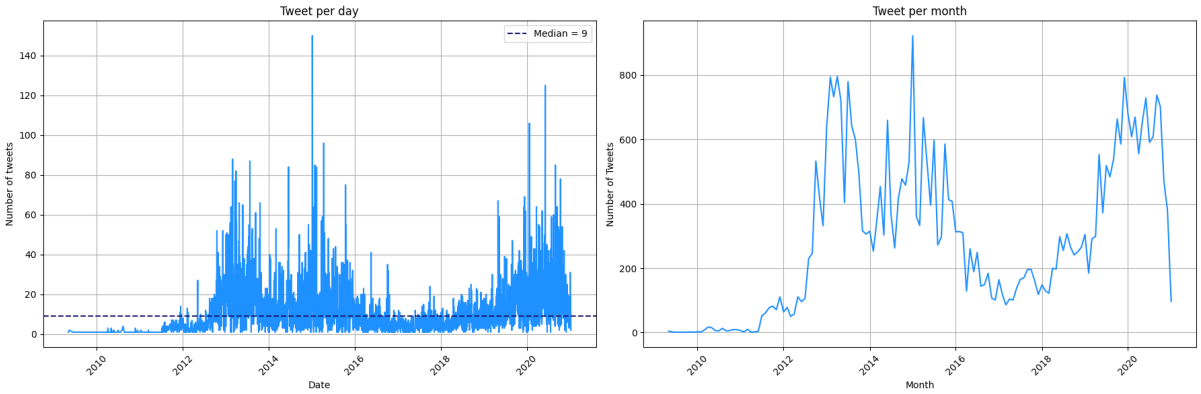

Figure 1: Tweet activity over time.

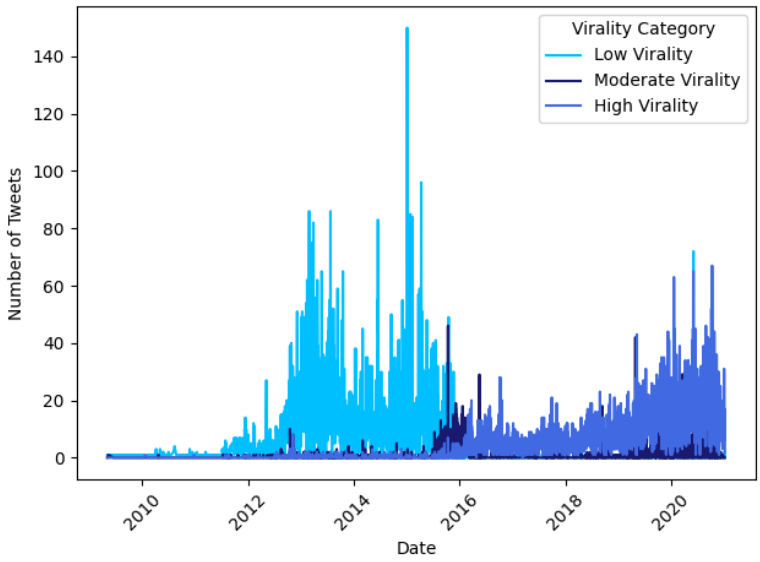

Figure 2: Virality trend over time.

Regarding sentiment distribution, the overall breakdown shows that the majority of tweets are labeled as positive, followed by neutral, with negative tweets being the least frequent. However, a closer look at the sentiment trends over time reveals more nuanced insights.

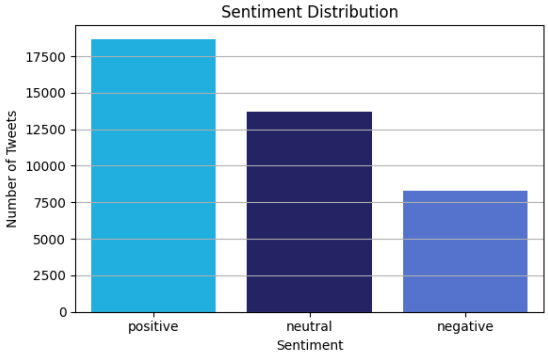

Figure 3: Sentiment Distribution

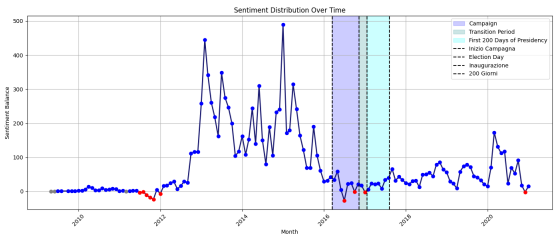

Figure 4: Sentiment Distribution Over Time

The temporal sentiment graph indicates that once Trump begins posting more actively, the overall sentiment remains overwhelmingly positive. As the 2016 election campaign draws closer, however, a noticeable shift emerges: the tone of his tweets becomes increasingly neutral or even negative. This trend stabilizes post-election, with only a minor uptick that does not match the intensity observed during the earlier campaign phase. This shift may signal a transformation in Trump's communication strategy across different phases of his public and political life on Twitter.

Election Impact on Tweet Sentiment

To determine whether there is a statistically significant difference in sentiment distribution before and after the 2016 election, a chi-squared test was performed, comparing sentiment frequencies prior to November 8, 2016, with those from that date onward.

A contingency table was constructed to represent the sentiment distributions in these two periods. The chi-squared test produced the following results:

- Chi-squared statistic: 1313.63
- Degrees of freedom: 2
- P-value: 5.62×10^{-286}

These results indicate a very strong statistical association between the election period and the sentiment expressed in Trump's tweets. The exceptionally high chi-squared value relative to the degrees of freedom, coupled with an effectively zero p-value, allows us to confidently reject the null hypothesis of independence. In other words, the sentiment distribution before and after the 2016 election differs in a way that is highly unlikely to be due to random variation, suggesting a meaningful shift in tone over time.

In comparing the sentiment distribution of Donald Trump's tweets before and after his election on November 8, 2016, we observe a notable shift in emotional tone. The proportion of positive tweets experienced a substantial decline, dropping from 53.56% before the election to 35.90% afterward—a decrease of nearly 17.66 percentage points. Conversely,

the negative sentiment category saw a modest increase, rising from 18.73% to 22.68%, a change of roughly 3.94 percentage points. However, the most significant transformation lies in the dramatic growth of neutral tweets, which expanded from 27.71% before the presidency to 41.43% after—a jump of over 13.71 percentage points.

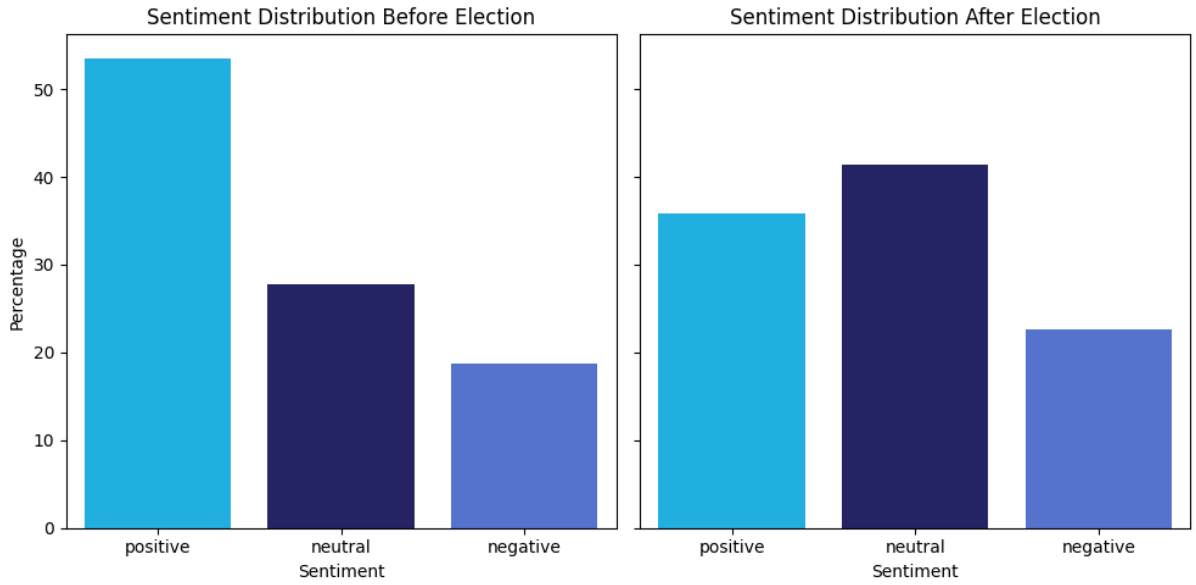

Figure 5: Sentiment Distribution Before and After Election.

While the rise in neutral tweets may suggest a shift toward a more restrained or ambiguous communicative style, it also introduces a challenge for sentiment analysis. Neutral content, by definition, lacks the polarity necessary to convey clear emotional stance, making it difficult to draw definitive conclusions about Trump's affective orientation post-election.

Beyond Neutrality: Emotional Trends in Post-Election Discourse

Upon inspecting the most frequent words in tweets classified as neutral—such as "president," "people," "democrats," "obama," "biden," and "election"—a pattern emerges that challenges the neutrality assigned by the model. Many of these terms frequently appear in politically charged or polemical contexts, often used with sarcasm, provocation, or implicit negativity.

Table 3: Top 20 Most Frequent Words

Word	Frequency	Word	Frequency
president	1609	years	418
people	785	going	416
democrats	692	today	415
obama	663	never	405
new	591	election	402
one	543	news	399
time	511	biden	397
like	492	vote	395
big	484	know	388
many	439	house	381

To address this limitation, I employed the third model trained on a richer taxonomy of 27 distinct emotions, supplemented by the neutral category. Furthermore, to avoid the potential dilution of interpretability caused by over-reliance on the ambiguous neutral class, I used the fourth model.

While the previous, coarser model revealed a significant post-election increase in tweets labeled as neutral, the third model confirms and deepens that insight. Specifically, tweets labeled as neutral rise from 37.91% before the election to 56.24% afterward—a shift of over 18 percentage points.

The detailed breakdown exposes a sharp decline in explicitly positive emotions. Emotions like admiration (–6.58), gratitude (–5.42), approval (–1.40), and caring (–1.28) all decreased significantly, contributing to an overall reduction in positive emotional tone of –19.17 percentage points. In contrast, the set of clearly negative emotions—especially disapproval, anger, and disgust—showed modest increases, amounting to an aggregate gain of less than 1 percentage point.

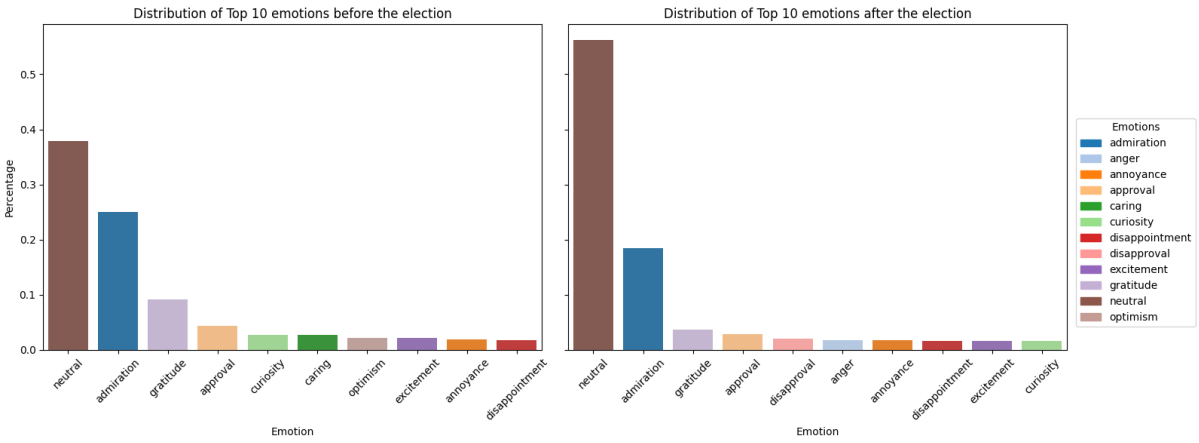

Figure 6: Distribution of Top 10 Emotions per Period.

This overwhelming dominance of the neutral class raises a fundamental concern: is this neutrality genuine, or does it instead reflect a limitation in the model's ability to detect

implicit emotional content?

Having excluded the neutral label in the fourth model, we are now able to observe the underlying emotional composition of what was previously a large and somewhat ambiguous category. The label neutral, which earlier masked a substantial portion of the data, is now disaggregated into specific emotions, revealing that a considerable majority of these tweets were in fact emotionally loaded, albeit subtly so. The most prevalent emotion replacing the neutral label is admiration, which alone accounts for 28% of the tweets previously marked as neutral. This is followed by approval (20.75%), annoyance (12.71%), and curiosity (6.89%).

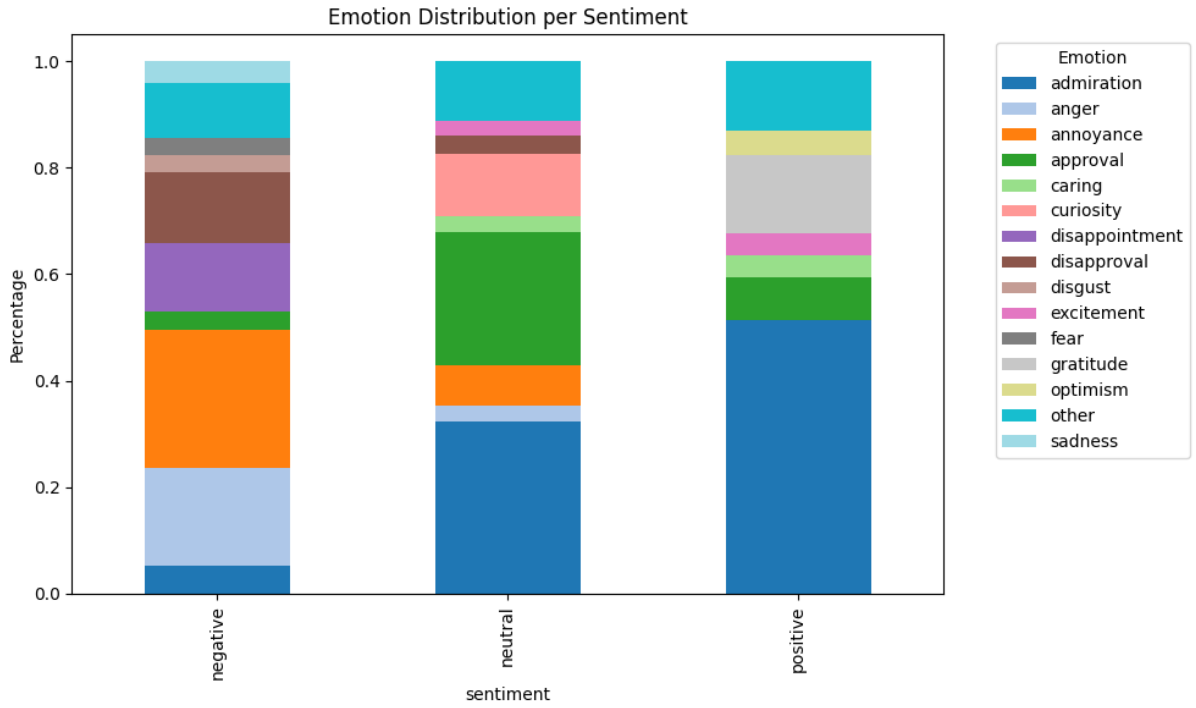

Figure 7: Sentiment investigation excluding the neutral category.

A particularly revealing insight emerges when examining the presence of retweets. Overall, retweets constitute only a small fraction of the dataset (roughly 17%). However, when narrowing the focus to tweets that were previously categorized as neutral but are now reclassified as admiration, the proportion of retweets jumps dramatically to 37.5%.

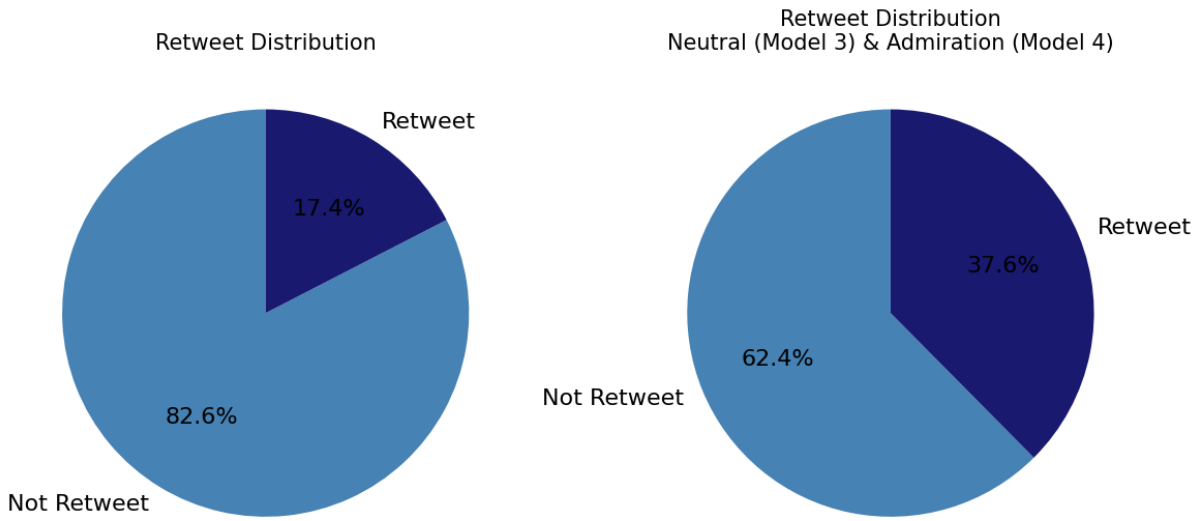

Figure 8: Retweets Distribution.

Turning to the temporal dimension of emotional expression, we observe notable shifts in emotional distribution before and after the election. Admiration remains the most dominant emotion across both periods, though it slightly decreases from 36.84% to 33.81%. More striking is the rise of approval and annoyance, which increase by over 4 percentage points each. Gratitude, on the other hand, experiences a sharp drop of more than 5.5 points.

Emotion	**Percentage Before the Election**	**Percentage After the Election**	**Percentage Change**
gratitude	9.18%	3.62%	-5.56%
annoyance	6.27%	10.58%	4.31%
approval	11.01%	15.10%	4.09%
admiration	36.84%	33.81%	-3.03%
anger	3.90%	6.55%	2.65%
disapproval	3.01%	5.40%	2.39%
caring	3.99%	2.09%	-1.90%
love	1.64%	0.45%	-1.19%
curiosity	5.83%	4.72%	-1.11%
excitement	3.10%	2.46%	-0.64%

Table 4: Emotion Distribution Before and After the Election

The significance of these changes is confirmed by a chi-squared test performed on the distribution of emotions before and after the election, yielding a test statistic of 1539.52 with 26 degrees of freedom and an extremely low p-value ($p < 1e$–308). This statistically significant result confirms that the emotional distribution shifted in a non-random way.

Manual Review of Emotional Intent in Trump's Tweets

One important consideration when analyzing the emotional tone of Trump's Twitter communication is the role of sarcasm and context-dependent meaning. While the models consistently suggest that positive emotions represent the majority of Trump's tweets, there remains a crucial ambiguity: some tweets use ostensibly positive language to convey a negative or provocative message. This concern is not new and has been explored in the paper "His Tweets Speak for Themselves: An Analysis of Donald Trump's Twitter Behavior" by Elayan, Sykora, and Jackson [2], which inspired a small-scale manual validation.

Previous research, such as [3], has demonstrated that combining contextualized embeddings like BERT with traditional word embeddings like GloVe can yield promising results in classifying sarcastic content—suggesting a valuable direction for future enhancement of emotion classification accuracy in politically charged discourse. Despite this potential, the present study does not incorporate automatic sarcasm detection techniques, adhering instead to the approach proposed by [2].

I selected a balanced sample of 200 tweets (36% negative, 33.5% positive, and 30.5% neutral) and manually reassessed the emotional intent behind each. After review, the new distribution revealed a notable increase in negative tweets (44%), with a corresponding drop in positives (27.5%) and a slight shift in neutrals (28.5%).

Several examples, analyzed using SHAP values [4], help to illustrate this ambiguity. In one tweet, Trump sarcastically remarks, "I am sure that Nancy Pelosi would be very happy to quickly work out free travel arrangements," seemingly offering help but actually mocking Democratic leadership. SHAP values show words like *happy* (+0.12) and *very* (+0.03) contributing to a positive prediction, despite the overall sarcastic intent. In another example, the tweet ends with "Welcome ISIS!"—where *welcome* alone contributes +0.37, misleadingly skewing the prediction toward positive, even though the message clearly conveys alarm or disapproval.

Example 1:
....It is done. These places need your help badly, you cannot leave fast enough. I am sure that Nancy Pelosi would be very happy to quickly work out free travel arrangements!

Example 2:
Remember when I said when Saddam Hussein fell, the new leader of Iraq will be meaner and tougher and hate the U.S. even more. Welcome ISIS!

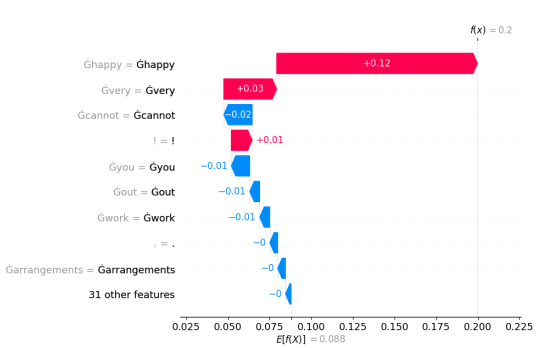

Figure 9: Shap Values for Example 1

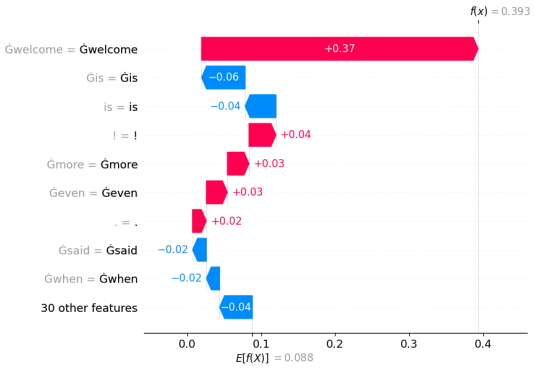

Figure 10: Shap Values for Example 2

Closing Remarks

This work set out to answer a central research question: is there a significant difference in the emotional tone of Trump's tweets before and after becoming president?

We observed a clear shift in the emotional landscape of his Twitter communication. Most notably, while positive emotions consistently represented the majority of tweets across models, there was a notable increase in negative emotions in the post-election period. The fourth model provided deeper insight by breaking down the "neutral" class into discrete emotions, revealing that admiration was the most prominent category and that emotions such as annoyance, disapproval, and anger became more pronounced after the election.

A manual review introduced an important disclaimer. It became clear that some tweets labeled as positive by the model were in fact sarcastic or contextually critical. This suggests that our models may overestimate the presence of genuinely positive sentiment.

Future work could benefit from more context-aware sentiment analysis methods that incorporate sarcasm detection, tone analysis, or even political relationship graphs to improve classification accuracy. Despite its limitations, this analysis offers valuable insights into the evolution and strategy of Trump's digital rhetoric.

References

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Guarav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[2] Suzanne Elayan, Martin Sykora, and Tom Jackson. His tweets speak for themselves: An analysis of donald trump's twitter behavior. *The International Journal of Interdisciplinary Civic and Political Studies*, 15(1):11–35, January 2020.

[3] Akshay Khatri and Pranav P. Sarcasm detection in tweets with bert and glove embeddings. In *Proceedings of the Second Workshop on Figurative Language Processing*, 2020.

[4] Scott M. Lundberg and Su-in Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.