The Role of Emotional Sentiment in Predicting the Popularity of Reddit Responses to Donald Trump's Presidential Legacy

R. Talvi, D. Vallmanya and I. Villalonga
Barcelona School of Economics

March 3, 2023

Contents

Abstract

The initial stage of our research involves accessing the Reddit API to extract all responses -including metadata- from the post "Donald Trump has not made a single lasting positive impact on the USA during his term as president". Subsequently, we undertake pre-processing activities such as handling missing values, stop word removal, tokenization, and lemmatization to construct the Document Term matrix. We then identify emotion analysis as a more suitable technique for forecasting document scores than sentiment and perform a descriptive analysis that relates emotion prevalence to document scores. Finally, we train a baseline model using top terms and compare it to our definitive model, which assesses the predictive power emotions have on document score or popularity.

1 Introduction

Reddit is a social news and discussion website where registered users can submit posts or links to various topics which are then given a score by other users. The most popular posts rise to the top while less popular ones are buried. Users can also comment on posts, creating a discussion thread that can be sorted by relevance or popularity. Posts are only open for a short time spam, which does not allow to perform relevant time series anlaysis.

In this term paper we will use Reddit's API to extract a huge number of users opinions in response to the prompt "Donald Trump has not made a single lasting positive impact on the USA during his term as president", which is included in a very popular reddit community called "Change my views". This reddit community underlying goal is to inspire controversial debates where users intervene attempting to convince interlocutors that have antoagonistic views.

There are many social networks that can be used as sources to construct a social thermometer of Trumps presidentail performance. Our approach is anchored on an analysis of a particular Reddit community that boasts a radical popularity, although we acknowledge that such a sample is not fully representative of the broader population. To that end, we suggest that future studies could potentially integrate an array of sources, such as Twitter, Facebook, polls, and other social media platforms, to more accurately capture the overall feeling of the public with respect to Trump's legacy. In this paper, however, our scope is much more specific and modest: we aim to discern the forecasting power that a document's emotional distribution has on its score and to determine which emotions are most influential in the evaluation and scoring process by reddit's users. Through our analysis of the relative importance of each emotion on document score, we aim to indirectly capture the dominant sentiment within the Reddit community towards Trump's presidency.

Our paper aims to answer the following concrete research questions: can we forecast the score of a document based on its prevalent emotions? What are the specific emotions that have explanatory power on the document score?

2 Web Scraping

In order to obtain the text bodies and metadata of the posts we want to analyze we make use of the Reddit API and PRAW, a Python library to facilitate information retrieval from the API. The Reddit API allows developers to access and retrieve data from the website, including posts, comments, user information, among other things. To make use of the Reddit API we use a popular library for accessing it called PRAW, "Python Reddit API Wrapper" that provides a simple interface for accessing and working with Reddit data.

In order to use PRAW developers need to create an account on Reddit and obtain their API credentials. Once the credentials are obtained PRAW can be installed and configured to connect to the API, allowing to scrape data such as post titles, comments, and other relevant metadata.

The Reddit API has rate limits that restrict the number of requests that can be made within a specific period of time and PRAW handles these limits by automatically retrying requests that fail due to rate limiting. For this reason our dataset only contains intially 725 observations.

3 Pre Processing

Data pre-processing is a crucial and essential stage before the data can be used to train models and perform text analysis. This stage involves cleaning, tokenizing, lemmatizing, removing punctuation, removing stopwords, implementing tf-idf and transforming the raw text data into a document term matrix that can be processed by any selected machine learning model.

3.1 Data set

Our dataset consists of 725 articles written by Reddit users in respons to the Promt "Donald Trump has not made a single lasting positive impact on the USA during his term as president". Below we display a sample of the original dataset.

	post_title	comment_awards	comment_author	body	comment_replies	comment_ups	comment_downs	comment_time	comment_score
0	CMV: Donald Trump has not made a single lastin...	1	unRealEyeable	How about his signing into law of "right-to-tr...	20	1646	0	2020-09-10 16:26:46	1646
1	CMV: Donald Trump has not made a single lastin...	1	optiongeek	I haven't seen this posted yet - but by far hi...	16	604	0	2020-09-10 18:16:28	604
2	CMV: Donald Trump has not made a single lastin...	3	Nateorade	Doubling the standard deduction on taxes is ab...	52	1654	0	2020-09-10 16:21:19	1654
3	CMV: Donald Trump has not made a single lastin...	13	NaN	I can't believe I'm doing this, but they say y...	101	6714	0	2020-09-10 16:25:47	6714
4	CMV: Donald Trump has not made a single lastin...	17	Ironclad_FTW	He lowered unemployment at record lows by the ...	12	1826	0	2020-09-10 23:17:06	1826

Figure 1: Sample of the original dataset.

Note that apart from the body column, which is where each article is located, we have the following metadata: "comment awards", "comment author", "comment replies", "comment ups", "comment downs", "comment time" and "comment score" (where comment refers to the article written by each Reddit user).

3.2 Towards Document Term Matrix

A document-term matrix (DTM) is a rectangular matrix that represents the frequency of words in a set of documents. The rows represent the documents and the columns represent the words.

Before constructing this matrix, which will later be analysed and processed by different models, several pre-processing techniques have been applied. First, we tokenize the texts by breaking them down into smaller units called tokens, which in our case are words. Once we have a bag of words we remove some stopwords (those that are uninformative or irrelevant for the analysis) and then we lemmatize them (by reducing them to their base or dictionary form). Later, we remove again the stopwords from the bag of words to make sure that the lemmatized version of all words is also removed. Finallly, we implement the Tf-idf technique which assigns a weight to each term in a document based on its frequency (TF) in the document and its rarity (IDF) across the entire corpus of documents (with the idea that terms that are more frequent in a document and less frequent in the corpus are more likely to be important and meaningful for the understanding of that document.) Then, we construct the document term matrix.

Note that in our preprocessing pipeline we chose to use lemmatization over stemming because it generally produces better results in terms of accuracy and relevance to the original text. This is because lemmatization considers the context and part of speech of the word, whereas stemming only removes the suffixes and prefixes of the word, which can lead to incorrect results in some cases. Also, we decide to implement

TF-IDF in our project because it helps to identify important and relevant words in a document based on their frequency and inverse document frequency. Finally, we also choose to add specific stopwords that are relevant to our particular subject, as these words only add noise in our analysis.

The following Figure shows some articles before and after applying the preprocessing. Note that in the column *body* there are the original (not preprocessed) texts and in the column *text preproc* we have the text preprocessed.

body	text_preproc
signing "right-to-try" legislation, allowing gravely ill patients access experimental drugs?	signing right try legislation allowing gravely ill patient access experimental drug
posted - important accomplishment keeping foreign conflicts. [According article] (https://www.theelders.org/news/only-us-president-who-didnt-wage-war), managed US foreign conflicts Jimmy Carter. Th...	posted important accomplishment keeping foreign conflict according article theelders org news didnt wage war managed US foreign conflict jimmy carter thus hold term second manage talk idiot danger...
Doubling standard deduction taxes absolutely net positive country. Previously rarely 12k line unless used mortgage interest deduction, itemizing unfairly balanced towards home owners (of previousl...	doubling standard deduction tax absolutely net positive country previously rarely 12k line unless used mortgage interest deduction itemizing unfairly balanced towards home owner previously benefit...
can't believe I'm this, argue favour things disagree sometimes..... said, easy. 1- He's donated entire presidential salary variety causes inauguration- VAs, education services plenty more. 2- conv...	believe argue favour disagree sometimes easy donated entire presidential salary variety cause inauguration VAs education service plenty convinced mexican government modernise labour part trade tre...

Figure 2: Example of some articles before and after preprocessing.

4 Sentiment Analysis

For the sentiment analysis we have used the *SpacyTextBlob* library from Python (combination of the *Spacy* and *TextBlob* libraries).

The sentiment analysis model in *TextBlob* is trained on a large corpus of text data to learn patterns in the relationships between words and their sentiment scores. Specifically, the model uses a lexicon-based approach that assigns scores to words based on their polarity (positive, negative, or neutral) and intensity (strong or weak). These scores are then combined to obtain a sentiment score for the entire text.

Once we obtained the sentiment for each article, we adjusted the scale to be from (0,1), so that 0 refers to negative sentiment and 1 to positive.

A couple of things need to be pointed out about the sentiment analysis results. We note that most of the texts are classified as neutral (sentiment very close to 0.5) and very few take values above 0.7 or below 0.3. We attribute this factor to the fact that many texts are short and without clear formatting making it difficult to capture sentiment well. Another fact we would like to highlight is that we expected to have a generally positive sentiment due to the fact that the prompt's formulation is possibly biased towards refuting the prompt.

In general, texts are classified as having either a very extreme (value close to 1 or 0) or neutral sentiment. We note that the classification is matched when the texts are radical (for positive or for negative). Otherwise, when it comes to sentiment-neutral texts, we see that most of them are quite positive.

Figure 3 shows some texts that are apparently positive but are classified with a neutral sentiment.

text	score	sentiment
can't believe I'm this, argue favour things disagree sometimes..... said, easy. 1- He's donated entire presidential salary variety causes inauguration- VAs, education services plenty more. 2- conv...	6714	0.572611
lowered unemployment record lows 2019. high percentage American's better financially president, according official studies. factually job openings unemployed. wages gone up. African-American unemp...	1826	0.509301
Doubling standard deduction taxes absolutely net positive country. Previously rarely 12k line unless used mortgage interest deduction, itemizing unfairly balanced towards home owners (of previousl...	1654	0.504346
signing "right-to-try" legislation, allowing gravely ill patients access experimental drugs?	1646	0.400000
fully funded Land Water Conservation Fund perpetuity, funded nearly necessary backlog repair work national parks. Great American Outdoors Act. more, believe disproves claim.	881	0.662500
posted - important accomplishment keeping foreign conflicts. [According article](https://www.theelders.org/news/only-us-president-who-didnt-wage-war), managed US foreign conflicts Jimmy Carter. Th...	604	0.448958

Figure 3: Example of texts with neutral sentiment.

We can see that the first row of the Figure shows the top scored text where the writter summarises all the positive things Trump has done in history. Nevertheless, it has assigned a neutral sentiment value . Then, we conclude that there is a lack of correspondence between sentiment and text.

In the next Figure we show the top articles with the highest and lowest sentiment.

text	sentiment
"Trump" written history books worst worst.	0.00
That’s rascist. Nasty, body shaming git shouldn’t he’s trash it. cheated votes	0.00
taken ultra-conservative racist movement brought forefront (I lie, worried) terrible undoing. hope.	0.00
retracted terrible “Dear Colleague” letter Obama admin place expelled university weakest accusation.	0.00
uae Israel alone insane event,	0.00
...	...
family member me, "He great economy virus came messed everything up"	0.90
opened eyes incredible amount racism country...	0.95
Best post 2020 period. done. Lock up.	1.00
pull couple thousand troops Syria. that's debatable whether ultimately best situation? dunno.	1.00
Best reason reason - started wars!!! Beat one!!!!	1.00

Figure 4: Top texts with highest/lowest sentiment value.

In this case, we can see a certain correspondance between the sentiment and text. As we have said, the extreme cases are correctly classified. The articles classified with negative sentiment values contain words as "worst", "racist", "terrible", while the positive senitment texts cointain "best", "opened eyes", "great economy".

Given the preceding analysis does not portray accurately the overall sentiment in a document, we will undertake an alternative approach that is closely related to sentiment analysis.

5 Emotion Analysis

In the realm of text analysis, sentiment analysis is a commonly used approach that aims to determine the subjectivity and polarity of a given text extract. Specifically, it assigns a probability to the text extract for having a predominantly positive or negative sentiment. In contrast, emotion analysis goes beyond polarity and attempts to identify and classify a range of emotions present in the text. This approach enables a more thorough understanding of the emotional content of the text and allows us to capture the complex interplay of various emotions that may be present in a given extract.

The NRCLex library is a Python package designed for sentiment analysis and emotion mining. It is a useful tool for extracting emotions and sentiment from text data. The NRCLex library uses a lexicon-based approach to analyze text. It comes with an emotion lexicon, which is a dictionary of words and their associated emotions. When analyzing a piece of text, the library searches for words in the lexicon and calculates their emotional intensity scores. These scores are then combined to generate an overall emotion score for the text.

5.1 Document and Emotion Assignment

In Figure 5 we present an illustrative example of how emotions can be assigned to a particular document. The figure displays the distribution of emotions across the document providing a visual representation of the emotions expressed in the text. We have selected the emotions that are more revelaing and likely prevalent in the context of our project: fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust and joy.

fear	anger	anticip	trust	surprise	positive	negative	sadness	disgust	joy	text
0.200000	0.200000	0.0	0.000000	0.000000	0.000000	0.200000	0.200000	0.200000	0.000000	signing "right-to-try" legislation, allowing gravely ill patients access experimental drugs?
0.058824	0.000000	0.0	0.176471	0.000000	0.294118	0.235294	0.000000	0.058824	0.058824	posted - important accomplishment keeping foreign conflicts. [According article] (https://www.theelders.org/news/only-us-president-who-didnt-wage-war), managed US foreign conflicts Jimmy Carter. Th...
0.153846	0.076923	0.0	0.000000	0.000000	0.384615	0.230769	0.153846	0.000000	0.000000	Doubling standard deduction taxes absolutely net positive country. Previously rarely 12k line unless used mortgage interest deduction, itemizing unfairly balanced towards home owners (of previousl...
0.074074	0.111111	0.0	0.222222	0.037037	0.185185	0.148148	0.037037	0.000000	0.111111	can't believe I'm this, argue favour things disagree sometimes..... said, easy. 1- He's donated entire presidential salary variety causes inauguration- VAs, education services plenty more. 2- conv...

Figure 5: Emotion proportion of a subset of texts

If we take a look on the second text it makes reference to Trump's administration success regarding foreign policy, namely, the almost unprecedented reluctance of America to militarily get involved in foreign conflicts. This results in emotions such as "trust" and "positive" being prevalent. However, the text also discusses a setback related to COVID, leading to the emergence of emotions such as "negative" and "disgust" in the emotional connotation. We explored other documents and found that the anlysis accurrately captures the prevailing emotion in almsot every document, although some emotions that receive a residual weight on the overall proportion are imprecisely assigned. The following document is representative and supports the preceding claim:

Document 9: sole positive turn blind eye China. motivations center, CCP tyrannical regime deprives basic liberty human rights systematically destroying economies countries dumping goods less cost destroy local industry addicted artificially cheap goods. It's opium. mention fentanyl export. regime bit deadly Hitler's Germany realization. mindset Chamberlain (just Europe away peace prosperity; vs let Hong Kong's freedom erode, Taiwan's claim independence unrecognized, away South China Sea). piss poor job alone approach. critical start standing China. It's pretty clear Trump's motives aren't centered though idolizes Putin Russia, subvert thuggish corrupt cleptocracy could. former leaders refused stand CCP. change Trumps presidency likely positive forward.

Emotion proportions:

fear	0.103448
anger	0.068966
anticip	0.000000
trust	0.103448
surprise	0.034483
positive	0.206897
negative	0.155172
sadness	0.086207
disgust	0.068966
joy	0.086207

Figure 6: Emotion proportion in a given document

The following excerpt discusses Trump's handling of the relationship with China -harsh and confrontational- citing it as an accomplishment due to China's dumping policies and status as a major source of drugs such as opium and fentanyl. As a result, trust and positivity prevail in the emotional tone of the text. In contrast, the document mentions Trump's close relationship with Putin, whose regime is characterized as a corrupt cleptocracy, eliciting negativity and fear, although to a lesser degree. Finally, there are a few instances of emotions, such as disgust and joy, that do not correspond with the text's content, though they are of little significance.

5.2 Top words in Emotions

In light of our preliminary analysis, where we identified "trust" as a predominant emotion, we sought to explore the specific words that are associated with this emotion. Given our initial uncertainty regarding the connotation of the word "trust", we constructed a word cloud to visualize the most frequent words that co-occur with "trust" in these documents. The size of each word in the cloud is proportional to its relative importance enabling us to identify the most salient terms that are associated with this emotion.

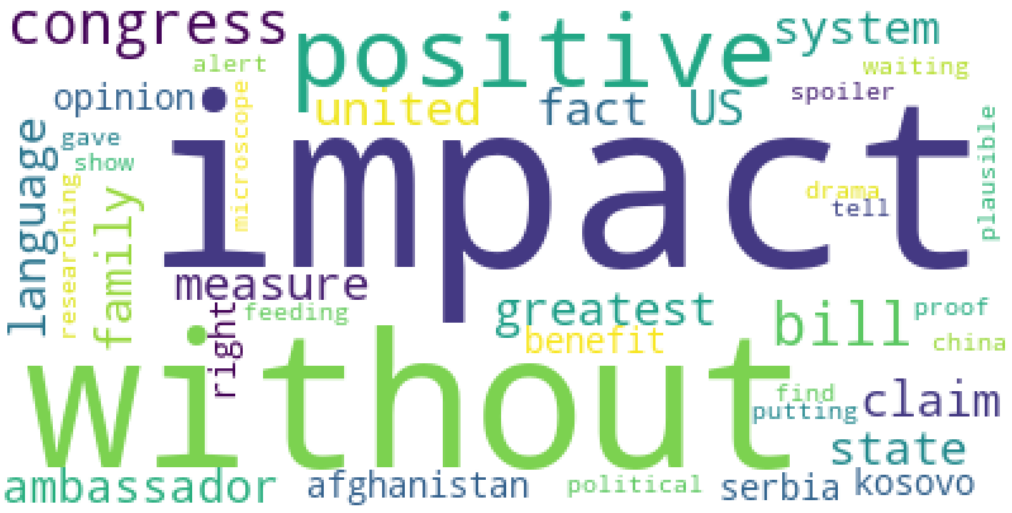


Figure 7: Top words in documents where Trust is the prevailing emotion

The word cloud highlights that the emotion of "trust" is closely associated with positive outcomes and perceptions. Interestingly, the neutral word "impact" is also among the most important words, in this context possibly reflecting positive achievements in responses about Trump's administration. The salient importance

of "without" along with the appearance of terms such as "Serbia" and "Afghanistan" suggests that the emotion of "trust" may be influenced by specific events or policies related to foreign affairs such as Trump's role during the negotiation of peace between Serbia and Kosovo and the full retrieval of US troops on Afghanistan during Trump's presidency respectively.

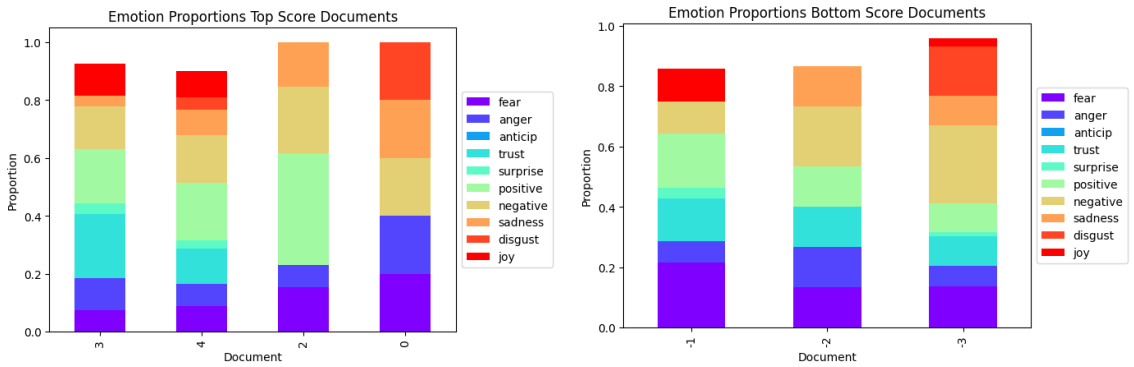
A word cloud for each emotion was done: we present two additional ones -for "surprise" and "negative" in the appendix 7.

6 Towards Random Forest Model: A Descriptive Analysis

As previously stated, the objective of our study is to investigate the relationship between the emotions conveyed in a given document and the score that users assign to it. Specifically, we seek to explore the predictive power of emotions in determining whether an document will receive a high or low score.
To investigate the predictive ability of emotions in determining the score of an article, we first performed a visual exploratory analysis to identify any patterns that may suggest a relationship between the two variables

6.1 Emotion proportion by Score

Firstly, we ordered the articles by their score and selected the top 4 and bottom 3 articles. Subsequently, we plotted the proportion of each emotion category for the selected articles, as shown in figure 8. This approach allowed us to gain insight into the emotional landscape of the top and bottom scoring articles, potentially indicating the presence of emotion-based predictors for article success.



(a) Emotion Proportions Top Score Documents.

(b) Emotion Proportions Bottom Score Documents.

Figure 8: Emotion Proportion for Top/ Bottom Score Documents.

Note that in the figure we can observe that not all proportions sum up to 1: this is because we have choosen to show a number of emotions and not all the possible ones given for each document.

Through our visual analysis of the emotional content in top and bottom scoring articles we observe a marked difference in the proportion of emotions. Specifically, we find that top scoring articles tend to have a lower proportion of fear and anger as compared to bottom scoring articles (represented by blue-purple tones). Conversely, positive and trust emotions are more prevalent in top scoring articles while these emotions do not appear to be dominant in bottom scoring articles. These findings suggest that articles with a positive emotional tone may be more appealing to users as indicated by their higher scores. We therefore hypothesize that users have an inclination to reward documents with a positive undertone.

6.2 Evolution of emotions proportion by Score

To gain a deeper understanding of the relationship between emotions and article scores we analyzed the evolution of emotion proportions across the top and bottom documents. Unlike the preceding analysis in figure 8, our goal now is to analyze independently the evolution of emotion proportions to seek if there is evidence an underlying pattern. Note that we focused on analyzing a limited number of documents - specifically, the top four and bottom four - given that considering a greater number of documents did not provide any additional relevant information.

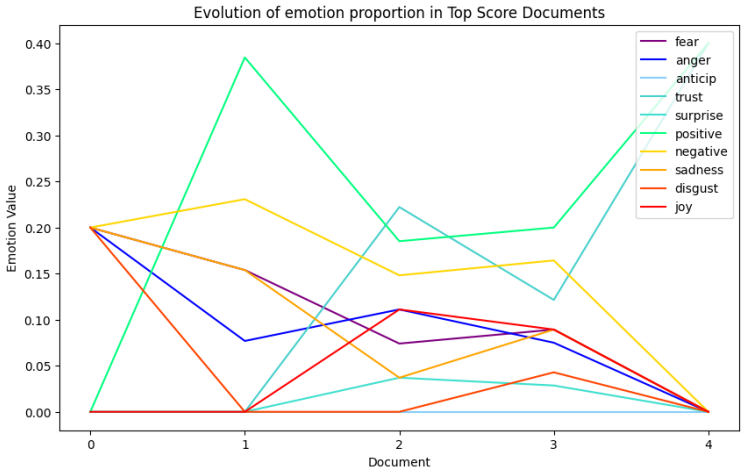


Figure 9: Evolution Emotion proportion Top Documnents

In examining the emotions exhibited in the top score documents we found an interesting pattern emerging. The relative distance between positive and negative emotions is consistently larger in favor of positive emotions, suggesting that users tend to gravitate towards texts that evoke positive feelings. Furthermore, trust is a recurring theme in almost all documents. In contrast, sadness and joy do not exhibit a clear prevalence of either emotion with their incidence rate interchangeably fluctuating among documents. Surprisingly, both sadness and surprise consistently exhibit low incidence rates indicating that these emotions may not be particularly impactful in determining the score of top score documents.

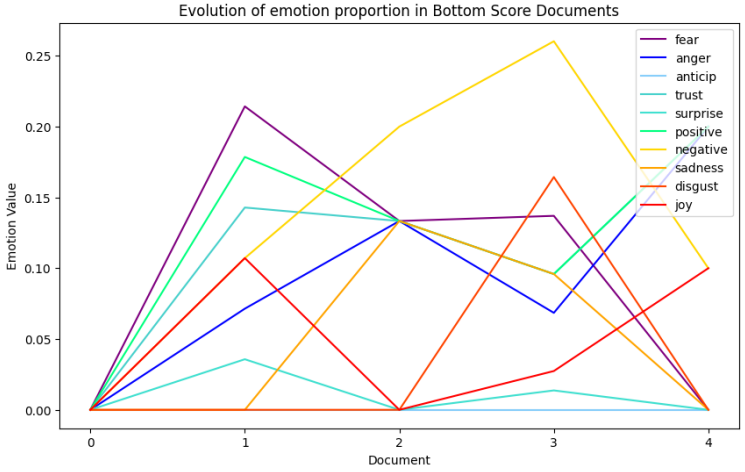


Figure 10: Evolution Emotion proportion Bottom Documnents

In contrast to top score documents, a different trend emerges when analyzing bottom score documents. The relative distance between positive and negative emotions is consistently biased towards negativity. Joy and sadness do not exhibit a clear relationship as they interchange across documents, suggesting an absence of stable patterns. However, in the documents where fear is present, its prevalence is consistently high. While anger is not present in all documents, when it is present, its incidence is high. Overall, these findings suggest that low scoring articles tend to show a prevalence of negative emotions.

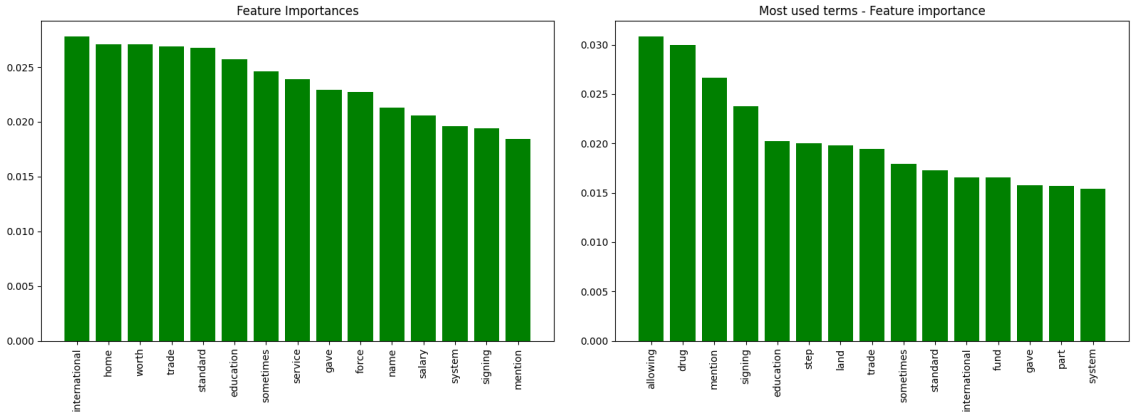
7 Random Forest Baseline Model

We are interested in seeing how the extracted features of emotions help us classify the texts in the post. In order to see whether we observe an effect we will compare a random forest regression using the emotional sentiment features compared to an analogous regression done using the most used 500 words in our corpus. Moreover, the emotional sentiment features extrated with the NCR library are obtained through a lexicon-based method, therefore using a most frequent words baseline that can help us detect if a particular word that signals an emotion is the one defining our results in the emotional random forest regression.

Choosing the most frequent words as a baseline model is also a convinient approach because it's simple and requires minimal feature engineering. It will provide a basic benchmark for comparison and can help to identify whether a more complex model is actually necessary. By comparing the performance of a more complex model to this baseline model, we can determine how much additional predictive power is gained from considering the emotional sentiment features. We will use the Mean Absolute Error metric to compare both models.

	MAE (train)	MAE (test)
Baseline RF model	25.78	33.52

Table 1: Mean Absolute Error (MAE) for the Baseline Random Forest model on the training and test set.



(a) Biased feature importances for most used terms.

(b) Unbiased feature importances for most used terms.

Figure 11: Biased (left) and unbiased (right) feature importances for most used terms

First thing to notice are some words that can tell us about possible discussion topics of Trump's presidency such as education, trade, international, or signing. We see that the feature importance values are not significant for this baseline model: we find no frequent word that defines a classification of the comment scores of our text bodies. We will see this better in the following section when we compare it with the performance and feature importances obtained with the model using the emotional sentiment features.

8 Random Forest Model

8.1 Training the Model

We propose to use a random forest model for our analysis. Random forests are a versatile machine learning model that can handle high dimensional data, provide quick predictions and are especially robust to outliers and non-linear data. Random Forest is chosen mainly beacasue our dataset contains several extreme values. Random forests are able to handle the outlier challenge by constructing multiple decision trees and averaging their predictions, resulting in a more robust and accurate model. Moreover, flexibility does not imply absence of interpretability as is often the case with machine learning models.

In order to develop a predictive model for the relationship between article emotions and user scores we first split our data into training and test sets. This is a common approach in machine learning to ensure that our model is capable of generalizing to unknown data. We then proceeded to train a random forest model on our training data using a parameter grid and cross-validation to tune hyperparameters and prevent overfitting. Specifically, we adjusted the parameters for minimum samples per leaf and maximum depth to optimize the model's ability to generalize to new data. This was crucial in avoiding overfitting as our evaluation performance metric was sensitive to random noise in the training set.

To evaluate the performance of our model we used mean absolute error (MAE) as the performance metric. The MAE measures the average absolute difference between the predicted score and the actual score, with lower values indicating better performance. In order to determine if our emotion classification adds any valuable information and increases the forecasting potential we compared the MAE of our model to a baseline model.

	MAE (train)	MAE (test)
Baseline RF model	25.78	33.52
Emotional RF model	19.52	24.44

Table 2: Comparison of Mean Absolute Error (MAE) for the Baseline and Emotional Random Forest (RF) model

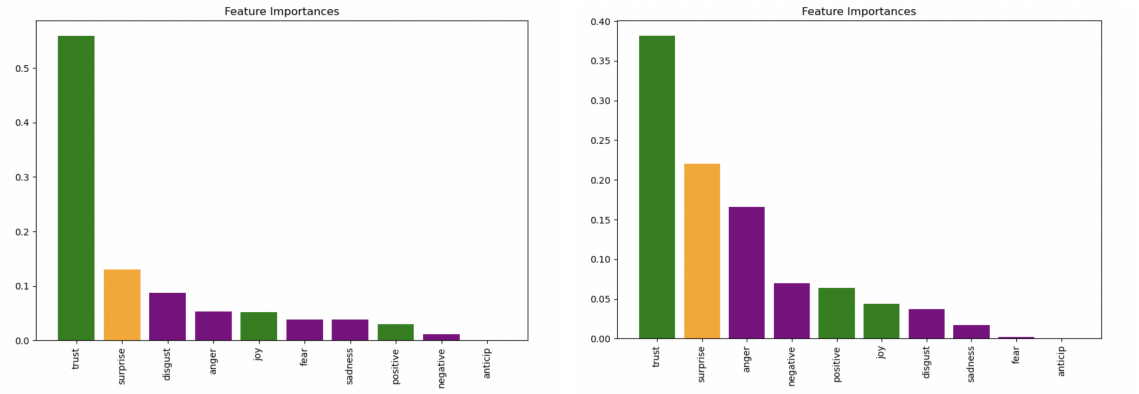
The scores in our dataset exhibit a wide range, from -16 to 6714. Despite this variability, a mean absolute error (MAE) of around 20 is considered relatively low for both the training and test sets. However, our analysis found that using emotional features resulted in a 35.6% and 37.1% increase in accuracy for the training and test sets, respectively, when classifying text body scores. Although the improvement in accuracy is only marginal, our model with emotional features outperformed the baseline model that exclusively used top words as features. Moreover, the model with emotional features demonstrated greater ability to generalize to new, unseen data. These findings suggest that emotional features were more effective than the baseline model and adds relevant additional information to forecast the document score.

8.2 Model Interpretability

Unlike some machine learning models -specially deep learning ones- random forest has the additional virtue of being efficient in forecasting without completely sacrificing interpretability. One of the main assets of Random Forest, beside virtues already commented, is that it provides a plot of the relative importance of each feature. Feature importance refers to a score that is assigned to each feature (i.e., input variable) indicating how useful it is in predicting the target variable. The importance of each feature is calculated by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest.

To address the bias issue associated with future importance in random forest models we can use a method known as permutation feature importance. This method involves measuring the importance of a feature by

calculating the increase in the model's prediction error after permuting the feature. If shuffling the values of a feature increases the model's error then the feature is considered important because the model relied on it for accurate predictions. Conversely, if shuffling the feature values does not increase the model error then the feature is considered not important because the model ignored it for the prediction.



(a) Random Forest Feature importance

(b) Random Forest Permutation Feature importance

Figure 12: Biased (left) and unbiased (right) feature importances

Firstly, it is important to note that we artificially grouped the emotions in our dataset into three clusters for the purpose of subsequent analysis. Specifically, we categorized emotions as positive (represented by green), negative (represented by purple), or neutral (represented by yellow). The positive emotion cluster included emotions such as trust, joy, and positivity, while the negative emotion cluster included emotions such as disgust, anger, fear, negativity and sadness. The surprise emotion was classified as neutral. This clustering decision was made to better understand the relative importance of positive and negative emotions in predicting document scores.

Comparing feature importance with and without permutation, we found that the permuted feature importance resulted in a more balanced representation of feature importance. In the biased representation, trust was the only feature that received significant weight, while in the unbiased representation, the feature importance was more balanced across emotions. In terms of ordering, the relative importance of negativity, which appeared negligible in the biased representation, gained some significance in the unbiased representation. Additionally, disgust, which was an important feature in the biased representation, received almost no weight in the unbiased representation. In contrast, positivity gained a relatively higher importance in the unbiased representation.

In our analysis we found that the emotion of trust, for which we previously identified the key terms in the word cloud, had the highest relative importance with a value of around 0.38. This suggests that trust has a strong relationship with the target variable and that its value significantly impacts the predicted outcome. Another important feature, with a weight of 0.22, was the neutral emotion of surprise, suggesting that documents containing unexpected content receive relatively more attention.

While trust, which is in the positive emotion cluster, was the most determining feature, negative emotions such as anger and negativity also had a role in explaining the association with the target variable with a combined weight of 0.24. Contrary to our initial suspicion based on exploratory analysis, emotions classified as strictly negative and positive received similar weights in the feature importance analysis. This suggests that positive emotions may not necessarily have a greater impact on predicting document scores, even though we found them to be more prevalent in top-scoring documents during the exploratory analysis.

9 Conclusion

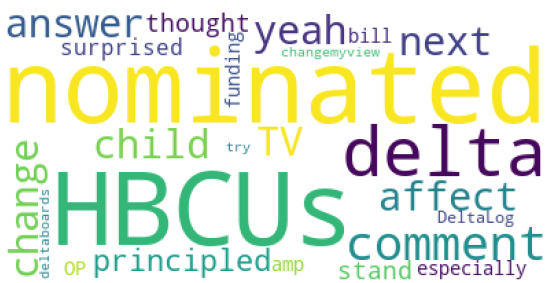
Our research aimed to investigate the role of emotional sentiment in predicting the popularity of Reddit responses to Donald Trump's presidential legacy. By extracting responses using the Reddit API and implementing data pre-processing techniques such as lemmatization and TF-IDF, we were able to construct a document-term matrix that succesfully provided the corpus for subsequent analysis.

Our approach to emotion analysis enabled us to identify and classify a range of emotions present in each document. We compared our definitive random forest model to a baseline model that used only the most frequent words in each document as features. Our approach demonstrated increased performance, particularly in its ability to generalize to unseen data.

Overall, our findings suggest that emotion prevalence has significant predictive power on the document score or popularity. Moreover, certain emotions, including trust, surprise, and anger, exhibited a higher degree of predictive ability in determining the score of Reddit responses regarding Donald Trump's presidential legacy.

Appendix

Emotion Analysis



(a) Words associated to emotion "surprise"

(b) Words associated to emotion "negative"

Figure 13: Emotions and word prevalence

LDA

We have implemented a Latent Dirichlet Allocation (LDA) model in order to extract the main topics of the Donald Trump articles. Note that we haven't obtained good defined topics due to the short lenght of the texts of our dataset.

Algorithm

Formally, LDA is a generative probabilistic model, in other words a model that attempts to explain how text documents are generated. The idea is that documents are represented as random mixtures over K latent topic, where each topic is characterized by a distribution over words. To set up the generative process we denote DirV (η) the Dirichlet Distribution with parameter vector $\eta = \eta_1, \eta_2, ..., \eta_V$ and $Dir_K(\alpha)$ is the Dirichlet Distribution of dimension K with parameter vector $(\alpha) = \alpha_1, \alpha_2, ..., \alpha_K$ The LDA assumes the following generative process for the corpus D:

1. For each topic k = 1,...,K:
   - Draw a distribution of words $\beta_K \sim Dir_V(\eta)$

2. For each document i = 1, ..., n
   - Draw a random vector of topic proportions $\theta \sim Dir_K(\alpha)$
   - For each word j = 1,...,d_i:
     (a) Draw a topic assignment $z_{ij} \sim$ Multinomial(θ_i)
     (b) Draw a word x_{ij} from p($z_{ij}|z_{ij}$,β) a multinomial probability conditioned on the topic z_{ij}.

Implementation

For the implementation of the LDA model we have used *gensim* an *nlk* modules in Python. By choosing the number of topics and number of words representing each topic we have obtained the following output.

For the model using prior we have set: 'war':0,'peace':0,'military':0, 'tax':1,'cut':1,'business':1, 'economy':2, 'employment':2, 'growth':2, 'unemployment':2, 'virus':3, 'pandemic':3, 'virus':3,'coronavirus':3, 'media':4, 'news':4.

Initially the project was planned to study the relationship between the sentiment of the documents given their proportion of each topic. But because the topics turned out to be general and not well defined, alternatives have been sought to carry out the project and LDA has not been used.

```
Perplexity: -5.83
Topic 0: ['right', 'republican', 'hate', 'white', 'work', 'racist', 'though', 'take']
Topic 1: ['country', 'china', 'change', 'bad', 'politics', 'action', 'political', 'power']
Topic 2: ['tax', 'american', 'job', 'administration', 'US', 'state', 'policy', 'cut']
Topic 3: ['positive', 'obama', 'impact', 'presidency', 'government', 'economy', 'best', 'real']
Topic 4: ['signed', 'bill', 'trade', 'order', 'reform', 'million', 'drug', 'funding']
Topic 5: ['war', 'peace', 'korea', 'israel', 'military', 'news', 'troop', 'UAE']
```

```
Perplexity: -4.90
Topic 0: ['right', 'country', 'positive', 'republican', 'government', 'take', 'hate', 'work']
Topic 1: ['china', 'change', 'bad', 'power', 'politics', 'action', 'political', 'talk']
Topic 2: ['tax', 'american', 'job', 'administration', 'state', 'policy', 'korea', 'unemployment']
Topic 3: ['war', 'obama', 'US', 'positive', 'peace', 'impact', 'israel', 'military']
Topic 4: ['signed', 'bill', 'trade', 'order', 'federal', 'reform', 'million', 'drug']
```

Figure 14: Visualizing topic proportions for a LDA model without priors (top) and with priors (bottom).

Logistic Regression

Logit Regression Results

Dep. Variable:	0	No. Observations:	575
Model:	Logit	Df Residuals:	568
Method:	MLE	Df Model:	6
Date:	Tue, 21 Feb 2023	Pseudo R-squ.:	-0.07297
Time:	19:34:14	Log-Likelihood:	-394.70
converged:	True	LL-Null:	-367.86
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
topic 1	-5.932e-05	0.000	-0.144	0.885	-0.001	0.001
topic 2	-0.3093	0.738	-0.419	0.675	-1.756	1.137
topic 3	0.2623	0.600	0.437	0.662	-0.914	1.438
topic 4	0.5725	0.723	0.792	0.429	-0.845	1.990
topic 5	0.3165	0.759	0.417	0.677	-1.172	1.805
topic 6	0.1413	0.584	0.242	0.809	-1.002	1.285
6	0.0298	0.588	0.051	0.960	-1.124	1.183

Pseudo R-squared value: -0.07296745865587217

Mean absolute error: 0.0812

Mean squared error: 0.0136

Figure 15: Logistic Regression Output