BUTTERFLY EFFECT

A data-driven peek at how people feel about the U.S. Republican Party and their candidate on Reddit ahead of the November 5th, 2024, presidential election

Javier Alejandro Borges Legrottaglie

University of Pisa

j.borges@studenti.unipi.it

Disclaimer. This work and its contents are for the purpose of model demonstration only and should in no way be seen or constitute an attack on any specific party, the data used for this work was sourced from the website www.Kaggle.com through the following link: https://shorturl.at/hsvC9. The comments within it were taken as if and without any modifications. Likewise, given the nature of Reddit's comment system, many comments may originate from the same user(s) as they may have been engaged in lengthy discussion / arguments with other users, potentially skewing the results in one particular direction.

Abstract. The following work is an attempt to understand the way users of the social media site Reddit feel about one of, if not, the most controversial political character of our time and the political party which he represents, i.e.: former U.S. president Donald Trump and the Republican party. In order to do that, I selected three tools for Natural Language Processing (i.e.: Sentiment analysis, Subjectivity analysis, and Emotion analysis) and applied them on a dataset containing comments on some of the communities present on the platform, i.e.: subreddits, to get a sense of what users are saying about this character ahead of the 2024 U.S. presidential election, an event that may actually impact, not only the American people, but the geopolitical environment as a whole.

Keywords: Politics, Geopolitics, Reddit, NLP, Machine Learning, Social media

1 Introduction

Not only has the aftermath of the global COVID-19 pandemic left most, if not all, countries in the world in economic turmoil, it has also exposed many issues with the current way that trade is being done nowadays. In particular, the fact that *Just-in-time* manufacturing on a global scale works well when there are no roadblocks in place as goods flow from one country to the next poses a problem if one link in the chain is unable to function properly, causing disruption in the markets which ultimately translates to higher prices for the every-day consumers.

Such difficulties became even more apparent after the Russian Federation's invasion of Ukraine, an event that caused shockwave around the world. The military campaign not only made Western-European countries reconsider both their stance and budgets on matters of security, but it also emboldened other political leaders around the world to lean in on their efforts with regards to territorial claims. Most worrying among them being the possibility of an invasion of the island of Taiwan by the People's Republic of China in an effort to enforce their "One China" policy.

Such events have also triggered an increasing amount of protectionist policies both in the European Union and in the United States during the last two presidential terms which signals what geopolitical expert and best-seller author on the subject, Peter Zeihan, described as "the end of globalization".

It is in this global context that the next American election is to take place, right in the midst of global changes that will make us challenge not only the way we see ourselves, but our place in this world as well.

As the election draws near, it has become clear that the 2024 presidential race will be between former president Donald J. Trump, and the incumbent Joseph Robinette Biden. Both of which are rather different when it comes to their political views. Former president Donald J. Trump has built a cult of personality that has disrupted American politics in a major way and has allowed controversial decisions to be made such as the overturning of Roe v. Wade, the building of a hugely unnecessary wall along the United States' southern border in an attempt to curb illegal migration and has also built a questionable rapport with totalitarian leaders, such as North Korea's Kim Jong Un, and Russia's Vladimir Putin.

While current president Joe Biden's administration and handling of many important issues has been far from perfect, there no arguing the fact a second Trump term may most likely cause irreparable damage to democracy in the United States, a further increase in isolationist and protectionist policies and generate distrust among the nation's allies in terms of trade and security, making a second Biden term preferable to many.

That is the reason why, in an attempt to provide some insight on how the election might turn out this November. I used a series of Natural Language Processing models and applied them to a dataset which contained posts and comments from the last 6 months from some key communities present on the social media site Reddit. The idea behind this is to find out how users of the site feel with regards to former president Donald J. Trump and the Republican party as a whole.

The following work is an explanation of the models used as well as an analysis of the results obtained from their application.

2 Methodology

As it was said previously, the objective of this paper is to figure out the way people talk about and feel with regards to former U.S. president, Donald J. Trump, and the Republican party as a whole, on the social media platform Reddit, whether users speak positively or negatively about them, the emotions people communicate in their comments and ultimately whether those comments are objective or subjective, i.e.: whether they express an opinion or actual facts.

To obtain this information, a dataset was extracted from the posts and comments on Reddit's main politics-related communities (subreddits) which spanned the last six or so months at the time of writing this paper, more specifically, from August 2023 to early March 2024. The content included within this file was then processed four different Natural Language Processing (NLP) models and subsequently analyzed using statistical tools to find out users of the site feel across the different subreddits present in the dataset.

3 Terminology

The foundation of the techniques used for the creation of this paper are listed below:

1. *Natural Language Processing*. *Natural Language Processing*, or *NLP*, is a field of artificial intelligence (AI) that deals with computers' ability to understand and process human language. *NLP* combines techniques from computer science and linguistics to enable machines to analyze, manipulate, and generate human languages like English. It uses machine learning and algorithms which are trained on massive amounts of data and to analyze the structure and meaning of text and speech and allows the identification of patterns and the understanding of the nuances of language.

2. *Sentiment Analysis*. *Sentiment Analysis,* sometimes referred to as opinion mining, is a task that falls under the umbrella of *NLP*. It utilizes such techniques to specifically target the emotional tone behind a piece of text. By analyzing the language used and looking for opinion words, users of this technique can determine whether the overall sentiment expressed in a piece of text is either positive, negative, or neutral.

   It does this by looking at words that have an intrinsic positive or negative emotion, e.g: awesome, terrible, cool, disgusting, etc., and assigns a polarity score to it, typically a number from (-1,1) where the closer it is to 1 the more positive it is, and the closer it is to -1 the more negative it is while the closer it is to 0, the more neutral the text is.

   This technique is really useful when trying to understand online store reviews as well as social media comments. However, it is important to note that different models function slightly differently and may be more or less useful based on the type of text, its content and last, but not least, its context.

3. *Emotion Analysis*. *Emotion analysis* is a more advanced field compared to sentiment analysis that delves deeper into the intricacies of human emotions expressed through text. While sentiment

analysis categorizes the general feeling as either positive, negative, or neutral, emotion analysis aims to identify the specific emotion of a text by giving it a score from (0,1) according to the degree of each joy, sadness, anger, fear, and even surprise the text conveys. The *NLP* algorithms do this by identifying words charged with emotional connotations, specific sentence structures, and even emojis.
Emotion analysis can be useful in various fields, such as with customer service interactions where understanding a customer's frustration can lead to better resolution, and social media monitoring where brands can estimate the way their audience feels with regards to a particular product.

Tools utilized.

In order to obtain the data necessary for my analysis, I exploited a series of different *NLP* libraries available for the Python programming language as well as *Matplotlib*, a visualization package used for the creation of the graphs shown on this paper.

First among them was *NLTK* which stands for *Natural Language ToolKit*, an open-source library which contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning and includes more than 50 corpora and lexical sources such as the *Penn Treebank Corpus*, *Open Multilingual Wordnet*, *Problem Report Corpus*, and *Lin's Dependency Thesaurus*.

Following this, I moved on to implement the first of the four models I chose to do the analysis with, i.e.: *TextBlob*, a *Python* library for *NLP* which uses *NLTK* to carry out its tasks. It is a simple library which supports complex analysis and operations on textual data.

It does this by assigning individual scores to all the words in the text and calculating its final sentiment by some pooling operation like taking an average of the overall text and returning the polarity and subjectivity scores of a sentence. *TextBlob* also includes semantic labels which help with its analysis by allowing it to assign meaning to *emoticons* (a sequence of characters meant to illustrate a facial expression), exclamation marks, *emojis* (a small image used alongside or in place of text), etc.

TextBlob is also able to assign the text a *subjectivity* score that lies between 0 and 1. It quantifies the amount of personal opinion and factual information present within the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

The second model used for this work was VADER (Valence Aware Dictionary for Sentiment Reasoning). a model used for *sentiment analysis* that is able to identify the polarity and intensity of emotions. It is available in the NLTK package and can be used directly on *unlabeled text data*, i.e.: it is able to determine such things.

VADER analysis relies on a dictionary that maps lexical features to emotion intensities known as *sentiment scores*. Such score of a text can be obtained by adding up the intensity of each word in the text.

Also, VADER is able to understand the basic context of these words, such as "did not hate" as a positive statement. It is also able to identify the emphasis of capitalization and punctuation, like "GREAT."

The end result of a sentiment analysis task through VADER yields a series of positive, neutral, and negative scores for each piece of text that illustrates the overall composition of the text, However, it also provides a compound score that ranges from (-1, 1) for the overall polarity of the text where -1 indicates its sentiment is totally negative, while a score of 1 indicates the opposite.

The third model employed for sentiment analysis was a distilled version of the Bidirectional Encoder Representations from Transformers (BERT) an open-source machine learning framework for NLP based on Transformers, which is a deep learning model where every output element is connected to every input element and renders calculations more dynamic.

What makes BERT special is the fact that it is able to deal with ambiguous language in text by reading the text from right to left and left to right at the same time, hence the name “Bidirectional”. However, the application of BERT is computationally expensive, that is why I opted for a distilled version of it.

In NLP, the term “distillation” refers to a technique where a model is reduced in size and complexity while losing hardly any language understanding capabilities. In the case of BERT, a distilled version of it named distilBERT is able to reduce its size and complexity by 40% while retaining roughly 97% of its capabilities.

For this work, I used a pre-trained version of the distilBERT model that had been previously fine-tuned for sentiment analysis on Reddit comments by Huggingface user Allison Gong (username: AG6019). This model produces a similar result to the ones previously mentioned, but it limits itself to providing positive or negative polarity scores, both are values ranging from (0,1) and represent how much of the text is positive and negative.

E.g.: The sentence “I don’t like pizza” yields the following results:

Negative	0.972
Positive	0.028

Table 1

The reason behind the decision to apply different models specifically for the task of sentiment analysis was to compare their results and how they interpret comments differently, additionally I was curious to see whether or not the distilBERT model was able to provide more interesting results that TextBlob and VADER so as to justify its runtime of +9 hours when compared to the previous two runtime of about 1 hour.

Last, but not least, I used a distilRoBERTa model for *emotion analysis* that had been fine-tuned for the English language by Huggingface user j-Hartmann, said model is a distilled version of the Robustly Optimized BERT Pretraining Approach (RoBERTa) model, a revisited version of the BERT model that is able to match and exceed the performance of its predecessor thanks to having better fitting hyperparameters and a larger training data size.

This model is able to identify the emotions present in the text, i.e.: anger, disgust, fear, joy, neutral, sadness, and surprise, and returning a value from (0,1) for each one of them.

E.g.: The sentence “This movie always makes me cry..” produces the following results:

Sadness	0.866
Neutral	0.047
Surprise	0.042
Disgust	0.015
Joy	0.014
Fear	0.009
Anger	0.007

Table 2

4 Statistical results

In this section, I will layout the composition of the dataset and the structures found within it.

The data used for this analysis was found in a .csv file relational table dataset structured in the following way:

Rows	1.297.856
Columns	24

Table 3

After a quick look at the data, I realized that the dataset contains records for over a million comments found within 23.447 posts across 30 different subreddits from the month of August 2023 to early March 2024, a more in-depth view of these distributions is found below.

N° of posts	23,447
N° of comments	1,297,856
Avg. posts per day	38.063
Avg. comments per day	1614.249

Table 4

Further analysis of the dataset revealed that, of the more than 1.2 million comments, more than half of them originated from the *r/politics* subreddit, while the rest of them came from other communities present on the site, namely *r/WhitePeopleTwitter*, *r/PoliticalHumor*, *r/VoteDEM*, *r/neoliberal*, *r/Conservative*, among others.

On the other hand, the number of posts per subreddit seems slightly more equilibrated, and while it is true that r/politics continues to dominate by having a little over 7,000 posts, this time it only represents roughly a third of all posts present in the data. Such difference indicates that the more 'specialized' communities have less user engagement per post. One could infer that such decrease occurs due to the fact that specialized communities often have less arguments in the comments since users who follow said communities tend to share the same political views.

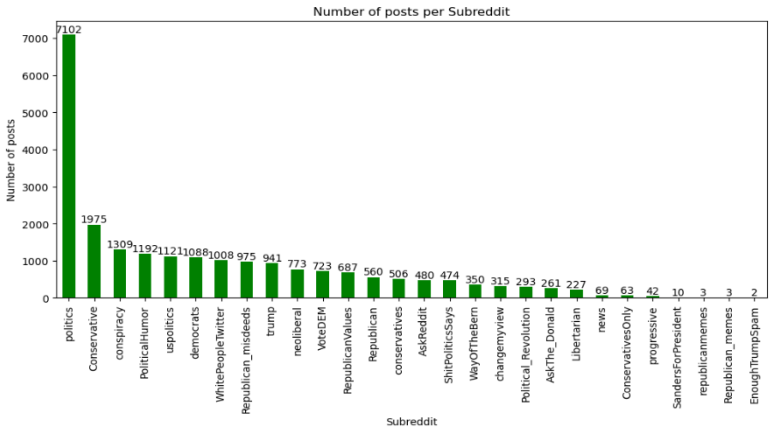

Figure 1.

The next step in the statistical analysis was to divide the dataset into two groups on which to do both the sentiment and emotion analysis.

The first group was formed by removing the rows which didn't include either the word 'republican', 'Republican', 'gop' or 'GOP', the last two being a widely used acronym used to refer to the Republican Party, it stands for "grand old party" and finds its origins in the early 1870s to emphasize its role in the preserving of the Union during the Civil War and its fight against the expansion of slavery into western territories, a position many would find ironic given the events of the last five years.

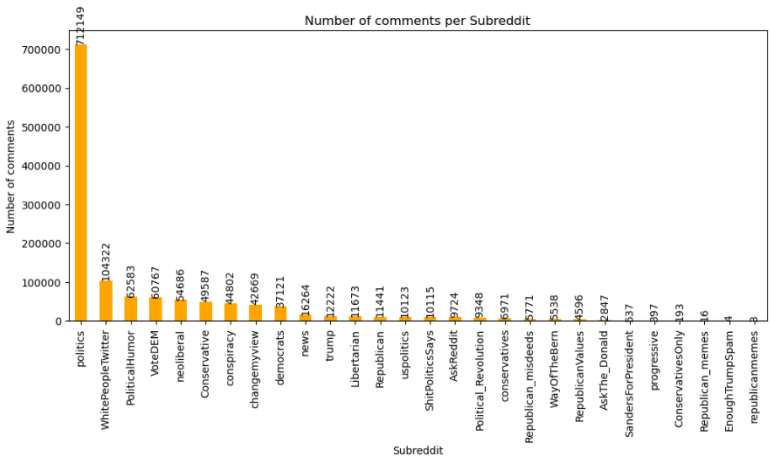

Figure 2.

The second group on the other hand, is made up of rows which contain the words 'trump' and 'Trump' in the text of the comments in order to see how people talk about the former president.

Name	N° of rows
GOP	122,082
trump	210,175

Table 5

As can be seen in table 5, the number of comments that explicitly mention the Republican Party are nearly 10% of the entire dataset while those mentioning trump account for slightly over two hundred thousand.

It is important to note that while it is true that this is a very small sample of all the comments available within the dataset, the aim of this work is to provide an estimate of how users of the platform feel about these two entities and while it would be interesting to get a more accurate rendition of this which could take into account entire comment threads that follow a sort of asynchronous conversation, it would ultimately be counter-productive.

Moving along with the partitioning, I figured it would be interesting to take a look at the different ways the sentiment towards former president Donald J. Trump and the GOP changes across the different subreddits. In order to do this, I subdivided both groups into 5 different groups comprised of different subreddits.

Table 6 reflects the subreddits chosen for each subset according to the way they align politically. Below is a brief description of the communities that comprise each subgroup.

Democrat	r/VoteDEM, r/democrats, r/RepublicanValues
Republican	r/Republican, r/Conservative
Politics	r/politics, r/ uspolitics
Neutral	r/WhitePeopleTwitter, r/news, r/AskReddit
Nutjobs	r/Libertarian, r/conspiracy, r/neoliberal

Table 6

- ***r/democrats.*** A serious subreddit dedicated to promoting the Democratic party by offering daily news updates, policy analysis, and opportunities to participate in the democratic process.
- ***r/VoteDEM.*** A subgroup of the *r/democrats* subreddit with the same values, although more geared towards activism and fundraising.
- ***r/RepublicanValues.*** A satirical subreddit where users share and discuss headlines about the modern Republican Party and its current woes.
- ***r/Republican.*** A subreddit dedicated to the GOP, self-described as a place for Republicans to discuss issues with other Republicans.
- ***r/Conservative.*** The self-described largest conservative subreddit. It is also a place where users post content that would bother the 'libs', i.e.: people who are liberal.
- r/politics. A subreddit for news and discussion about US politics in which they share articles from many different news sources that lean both left and right.
- ***r/uspolitics.*** A subreddit for US political discussion, a way smaller subset of *r/politics* where the content's level of granularity is more US centric and less global.
- ***r/WhitePeopleTwitter.*** A subreddit with no political affiliation in which people share memes and jokes they find on social media site X, formerly known as Twitter. The name of the subreddit refers to the fact that, on average, white people in the US have different backgrounds and therefore humor when compared to the content ported on r/BlackPeopleTwitter.
- ***r/news.*** A subreddit where news articles about current events in the US and the rest of the world are shared and discussed.

- ***r/AskReddit.*** The subreddit where users go to ask and answer, as they define it, thought-provoking questions.
- ***r/conspiracy.*** A subreddit where users share their thoughts and discuss conspiracy theories. The community describes itself as a place to 'keep an open mind', but its content is limited to discussions of people who are either famous or in politics and the possible links and stories associated to them.
- ***r/Libertarian.*** A subreddit where users the ideology of the same name as well as share and discuss news related to it, likewise its description explicitly says it is forbidden to share content related to both Communism and Socialism since they are antithetical to libertarianism.
- ***r/neoliberal.*** A Reddit community where users discuss news about trade regulations, the description of the subreddit states that their values are "Free trade, open borders, taco trucks on every corner".

After having subdivided the GOP and Trump groups into the 5 subsets described in table N, I went on to extracting the general sentiment for each comment as well as their emotions by applying the models previously mentioned.

5 Sentiment analysis and model comparison

In this section, I will lay out the results obtained from the application of the models mentioned in section N, look at the way the scores are distributed across the data as well as providing some examples of the comments each model identified as the most positive and negative.

5.1. VADER results.

Applying the VADER model yielded the aforementioned scores for each comment, i.e.: the positive, neutral, negative, and compound. I chose to focus on the compound score for each comment and take a look at what the overall score is for each subset for both the subsets that contain comments mentioning the former US president, as well as the sentiment regarding the GOP itself.

It is important to note that for the application on this model, the version of the VADER lexicon package used for this work was that which was available before March 12, 2024. \

Table 7 shows two examples of the results obtained from its application.

Comments	Results
If I could make social media disappear, Tik Tok would be the first to go.'	Negative: 0.128 Neutral: 0.872 Positive: 0.0 Compound: -0.2263
Idiots comforting idiots	Negative: 0.000 Neutral: 0.426 Positive: 0.574 Compound: 0.4019

Table 7

Trump subsets.

When looking at the resulting compound scores as reported in figure N, one can see that there is a lot of variation among the scores.

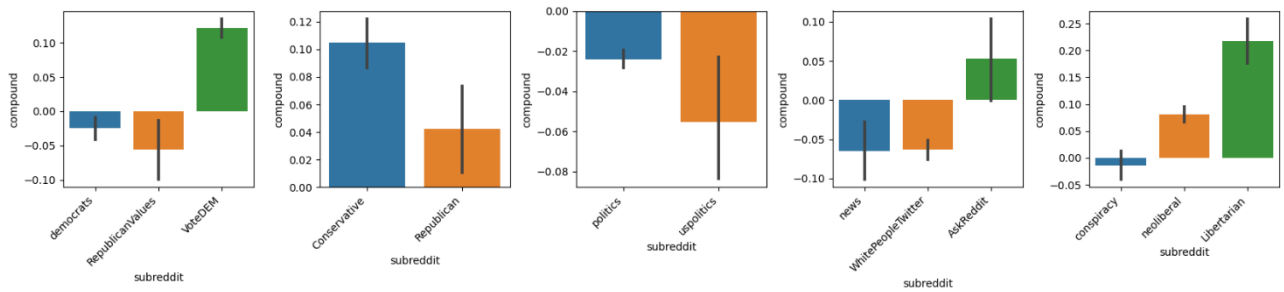

Figure 3 Compound scores for the Trump group of comments

For instance, the democrats subset yielded contrasting results, where the VoteDEM subreddit has an overall positive score with little variance (as shown in the error bar) which one would not expect, while the democrats and RepublicanValues subs yielded more predictable results given their nature, i.e.: the first one being more professional while the latter being more satirical.

Moving on, looking at the neutral subset reveals an overall negative score, but with a high variance, meaning that the sentiment expressed in the comments of these sections are actually and by all means neutral. The same occurs for the political subset, in which the sentiment is just barely negative.

And lastly, the republican and nutjobs subsets show interesting results, while most people would have guessed users who comment on *r/Conservative*, *r/neoliberal*, and *r/Libertarian* to speak about the former president in a more positive way, one can't help to notice the fact that comments from *r/Republican* remain neutral and with a high variance which seems to indicate that there are conflicting opinions about Mr. Trump in that reddit community.

Here are some examples of highly negative and positive comments as scored by the VADER model:

Comment	Score	Subreddit
I don\'t know why I\'m still taken aback at the level of cowardice at play, standing by and allowing Israel to ethnically cleanse its Palestinian populous. Every few weeks we learn that "privately Biden can\'t stand Netanyahu" or that in the strongest possible terms he\'s asked nicely for them to stop committing crimes against humanity. It\'s the weakest possible shit, to claim to have principles in private but to display no backbone on them in public. It\'s fucked that the price of preventing a second Trump presidency is that a lot of brown people *have* to die half a world away. It\'s like a blood sacrifice for democracy. It\'s so cool that the machinery of which Biden is ostensibly in charge cannot be prevented from facilitating the bulldozing of Palestinian neighborhoods, the starving of Palestinians, the murder of children. Gotta kill brown people or no more democracy. Shit, are people noticing that we kill a lot of brown people? Shit shit shit some of those brown people have families who vote. In AMERICA?! When did that happen? Oh fuck and everyone under the age of like 35 agrees that we	-0.9983	politics

shouldn't be engaged in a genocide. That\'s so many voters. Ah shit guys... I uhhh... I think we still have to keep killing brown people, no way out of it. Yeah, I know. Bad news, Trump might be president because "killing brown people" ranks really high on the list of things we cannot stop doing under any circumstances.		
Upon complete physical examination of Donald J. Trump, we have found: He stands 7 feet tall and possesses a truly grand stature (breathtaking). He easily maintains the perfect physique at all times (no one on our staff has ever seen a better physique!!!). And he maintains an incredible level of bravery and stamina (scientifically comparable to a full grown Male Kodiak Bear, but even more strong and very very brave!!!). It was also observed that he emits a faint glowing aura, which in our unbiased scientific opinion confirms he is always winning and the most perfect (so true!!!!). In closing, it's clear with this perfect very very strong and brave physique, he could easily live to the age of 300 years and beyond. Our entire staff agree that everybody knows this.	0.9975	conspiracy

Table 8

GOP subsets.

As for the results of VADER on the GOP subset, we get similar results to the Trump subset, but with some key differences.

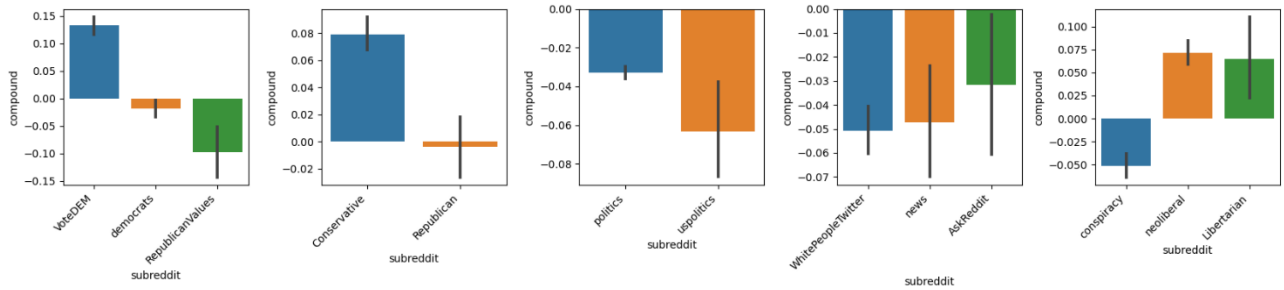

Figure 4 Compound scores for the GOP subsets

First of all, as can be seen in figure 4, we get different degrees of sentiment with regards to the GOP in the nutjobs subset, from the neutral result of *r/conspiracy* to an average score of 0.09 in *r/neoliberal* and a very positive 0.22 for *r/Libertarian*. This last one makes sense since the Republican Party has been against public spending and economic safety nets in recent years.

Here are some examples from these three communities:

Comment	Score	Subreddit
For republican politicians?? HA HA HA

Trump is a loser. Those supporting him right now have crafted a fiction that does not match up to reality. 4 time loser who failed to implement any lasting change his first term. You can screech about campaigns all you want. Trump's only saving grace was being indicted. Being a victim of the left. How pathetic is that? There was no donors dictating, your fantasy obsession doesn't get rewrite history. DeSantis bucked two of the biggest GOP donors in Florida. AB and Disney. The Trumps sucked up to both. Do you think the donors were scrambling for a 6 week abortion ban? No, Trump cowardly ran from a strong conservative stance. As he does everything. Martyrdom is what carried the corpse of Trump's campaign to the surface. No one anywhere has a political strategy that deals with that. As it stands they are executing the obvious. DeSantis Delivers. He has the record and the plan. Trump campaign lies. Over and over and over again. Be it fake gun buys. Fake union endorsements. Fake claims about Florida. Fake claims about Casey DeSantis' cancer. It's fake to it's core. No authenticity. But the worse part is, he's a loser. I could ignore his failings if he could actually win. He can't. He's done next year with 0 path to the presidency. He's busy grifting you to pay his legal fees while he is set to land multiple felony convictions and is going to take a landslide loss. But go ahead and support the loser. He failed on America First his first term. Be failed on Stopping the Steal after sucking up 250 million in donations. He failed with his candidates in 2018, 2020, 2022, and 2023. His speaker McCarthy failed. His RNC chair woman he had installed has failed.	-0.9975	Conservative

Table 9

5.2. Text Blob results.

When it came to the Text Blob model, I took the scores for each post that ranged from (-1, 1) and created a labeled the comment as negative if the score was below 0, neutral if it was exactly 0, and positive if it was higher than 0, not that these labels do not reflect the degree or positivity or negativity associated with the comments, so the following are not to be interpreted as “too positive” or “too negative”, they just indicate the general sentiment in the comments.

Likewise, Text Blob also provides a Subjectivity score, i.e.: a value from (0,1) that measures how much subjectivity there is the higher the value. For the purposes of this work, I decided to label the comments with a value greater than 0.33 as *subjective* while labeling those with a value lower than 0.33 as *objective*.

Examples of the results obtained from this analysis can be found in the following table:

Comment	Result
If I could make social media disappear, Tik Tok would be the first to go.	Polarity: 0.141667 Subjectivity: 0.20 Label: Objective

Idiots comforting idiots	Polarity: -0.80 Subjectivity: 0.80 Label: Subjective

Table 10

Trump subsets.

A quick glance at the results reveals that the results are quite different to what was produced by the VADER model. In particular, one can see that the democrats subset gave contrasting results to the previous model since it assigned a positive score to a stark majority of the comments present there, which seems counter intuitive since one would not expect users who comment on such subreddits to be particularly favorable to the former US president.

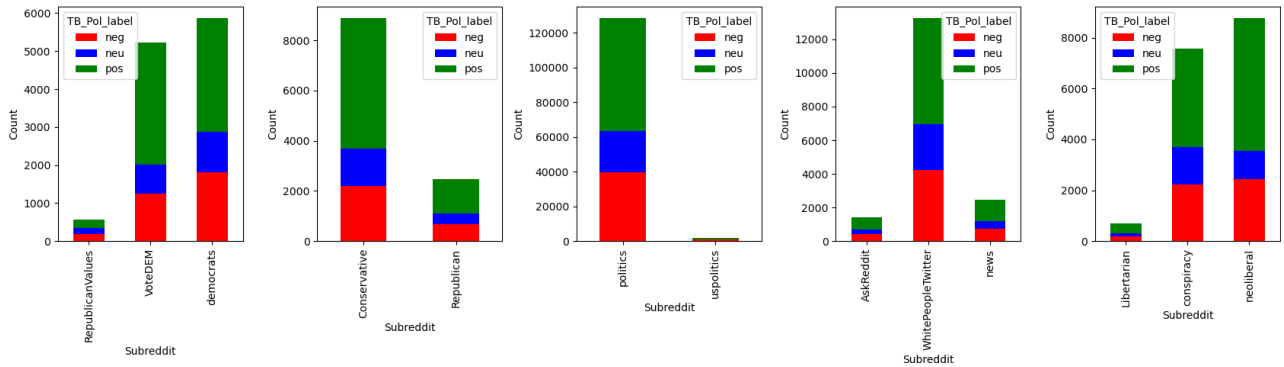

Figure 5 Text Blob polarity scores for the Trump subsets

Furthermore, Text Blob also assigns a positive label to a majority of the comments across all subsets, leading to little-to-no information or insights gained.

In terms of objectivity, we can see in figure N that a rough average of 70% of comments present in all subsets show a subjective opinion, this is to be expected since sharing objective information in a comment section of a social media platform is the exception rather than the rule.

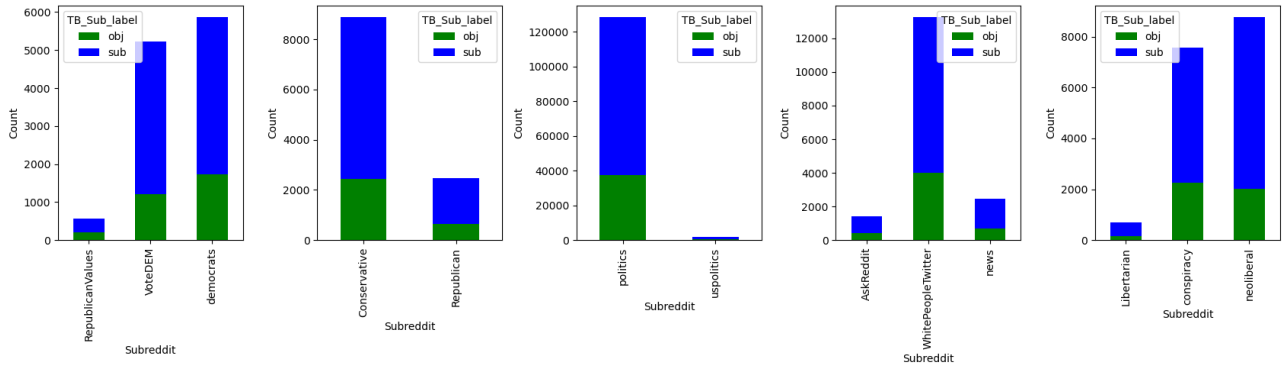

Figure 6 Subjectivity scores for the Trump subsets

Below are some examples of the comments the model identified as very positive, very negative, very objective, and very subjective:

Comment	Polarity score	Subjectivity	Subreddit
Yes, but it pisses Trump off, so it's a win!	1.0	0.4	democrats
Although all the cases against Trump have elements of fakeness and weakness, this one might be the worst.	-1.0	1.0	Republican
Awesome. Let them have what they want. Bankruptcy … just like the Trump companies.	1.0	1.0	conspiracy

Table 11

GOP results.

Moving on to the GOP subsets, figure 7 shows similar results to those in the Trump subset, i.e.: mostly positive comments across all subreddits, sadly there is little insight to be found here.

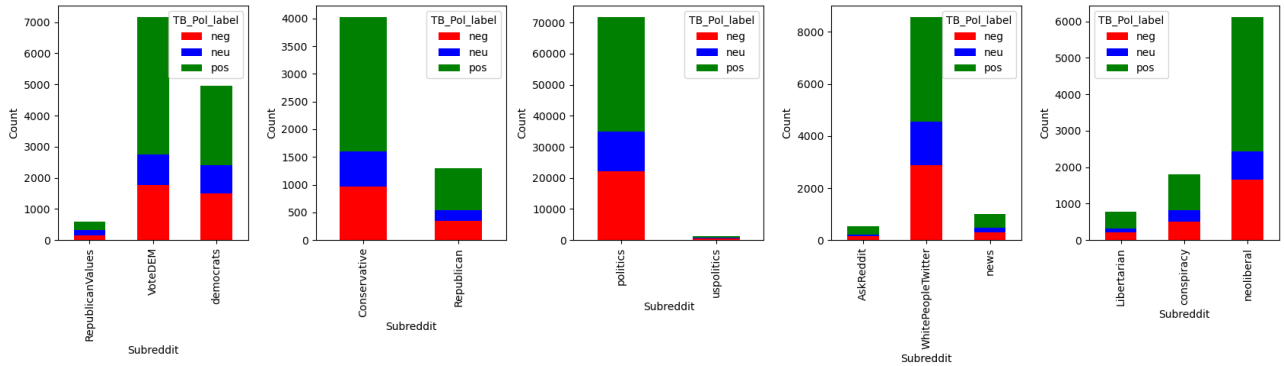

Figure 7 Text Blob polarity scores for the GOP subsets

In terms of objectivity, very little difference in to be found, if anything, we can point out that there is a drop in the amount of objectivity present in the *r/conspiracy* as well as in the *r/democrats* subreddits. Other than that the results are pretty much the same.

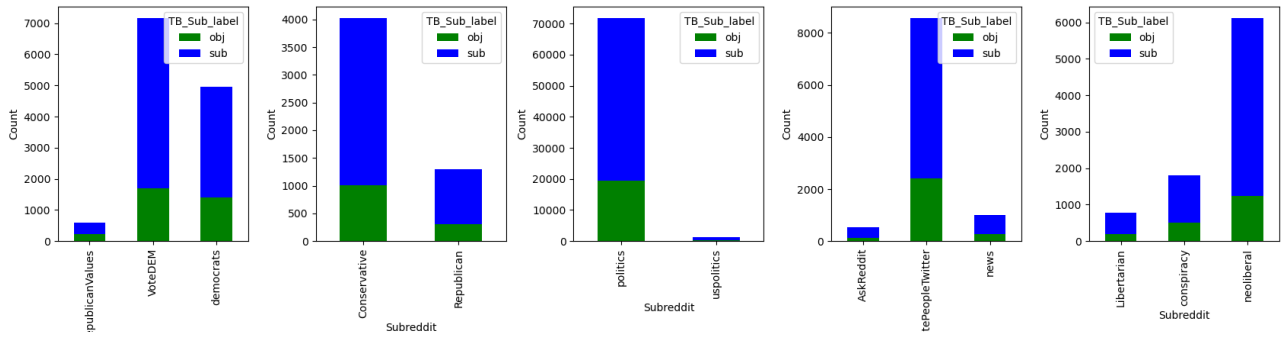

Figure 8 Subjectivity scores for the GOP subsets

Below are some examples of the comments Text Blob identified as very positive, very negative, very objective, and very subjective for this subset:

Comment	Polarity score	Subjectivity	Subreddit
Evil Republicans people need that money!!! 🤣 /s	-1.0	1.0	Conservative
He tried this bipartisan best buds approach for years, and Republicans immediately moved to impeach him and continue to bitch about voter fraud. Yeah nah.	1.0	0.3	neoliberal

Table 12

5.3. distilBERT results.

Last, but not least, the results of the distilBERT model should be regarded as the most accurate of the three models applied, since it was trained using Reddit comments themselves and it was specifically fine-tuned for this task. Once again, it is important to consider that the distilBERT model returns two scores that range from (0,1) that indicate how much positivity and negativity there is in each comment, and for the purposes of simplicity, a decision was made to provide a single label for each comment according to which score was higher for that particular comment.

e.g.: If a comment had a negativity score of 0.49, and a positivity score of 0.51, then that comment would be labeled as 'Positive'.

Examples of comments evaluated by this model can be seen in the table below:

Comment	Results
If I could make social media disappear, Tik Tok would be the first to go.	Negativity: 0.017772 Positivity: 0.982228

Idiots comforting idiots	Negativity: 0.980899 Positivity: 0.019101

Table 13

Let it be noted that, even though the first comment speaks ill about the social media platform Tik Tok, the words used in that comment give it an overall positive score while the second comment feature a pejorative adjective which radically increases the amount of negativity present in the comment.

Trump subsets.

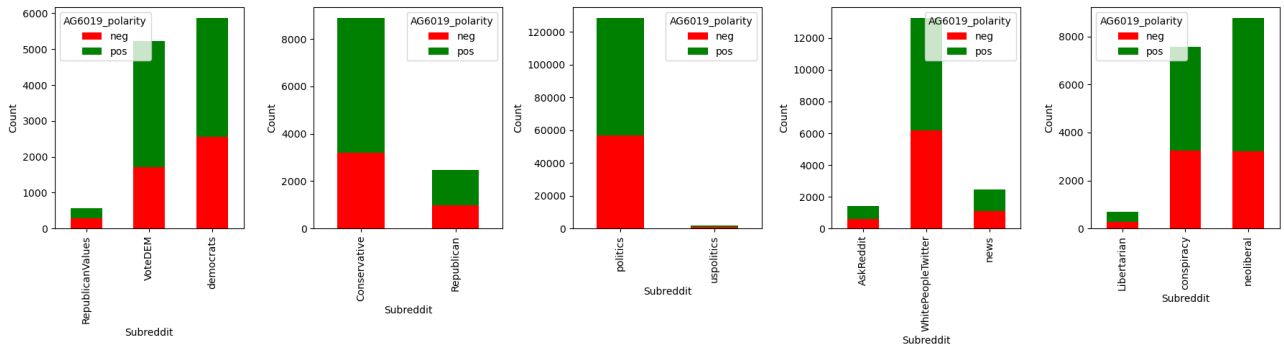

Figure 9 distilBERT polarity scores for the Trump subsets

As can be seen in figure N, results here are quite similar to those provided by the Text Blob model in the fact that it shows that the majority of comments show a positive sentiment in the majority of the comments. However, it is important to note that distilBERT is unable to tell certain subtleties in the language and will classify an otherwise neutral comment as positive if the language found in the text is not intrinsically negative, i.e.: pejoratives and so on, as can be seen in the examples provided in table 14.

This rational would provide an explanation to the unexpected number of positive comments present in the democrats, politics, and neutral subsets.

Below are some examples from this group of comments.

Comment	Negativity score	Positivity score	Subreddit
What a disgusting political hit job. Could you imagine how Trumpworld would have reacted if Mueller testified as a private citizen? Shame on you Robert Hur, you sick fuck	0.9865415	0.013458497	neoliberal
Perfectly fair my man and I agree with you - I'm a Desantis guy, but the Trumpism is just too popular right now	0.0051171076	0.99488294	Conservative

Table 14

GOP subsets.

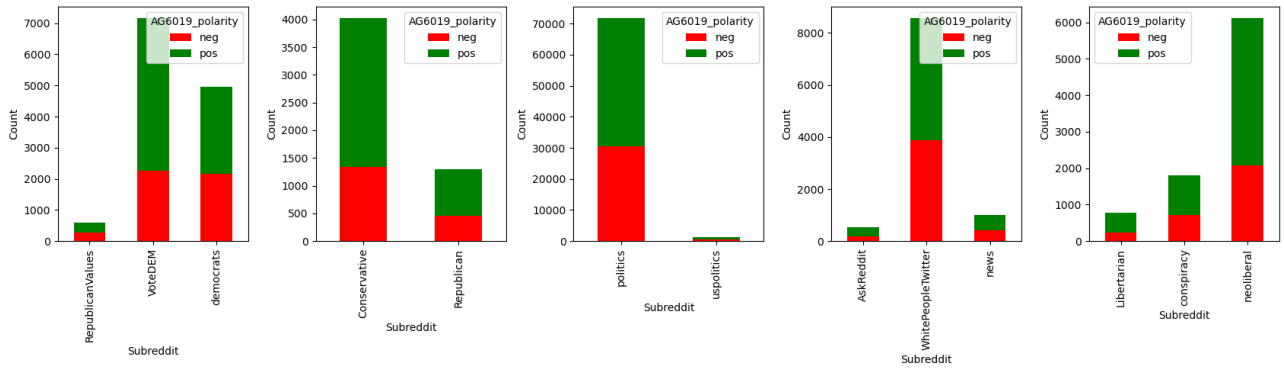

Figure 10 distilBERT polarity scores for the GOP subsets

Moving on to the comments mentioning the Republican Party, figure shows that sentiment towards the GOP across these communities is relatively more positive when compared with those mentioning former president Donald Trump. In particular it can be seen that *r/conservative*, *r/voteDEM*, and *r/neoliberal* share more or less the same percentage of negative/positive comments. However, it must be kept in mind that distilBERT labels comments that one would consider neutral as non-negative, i.e.: positive.

Below are some examples of the results found in these subsets.

Comment	Negativity score	Positivity score	Subreddit
Yet one more real world example of why voting for Republicans is voting against your own best self interests.	0.005215484	0.99478453	democrats
remember when Obama spent years being Deporter in Chief for political capital and Republicans told him to go fuck himself? how are we falling for this shit again? they don't give a fuck about security they *just* hate immigrants	0.9863929	0.01360703	neoliberal

Table 15

6 Emotion analysis results

As it was previously reported, the distilRoBERTa model used for this analysis was fine-tuned for the detection of emotions in text written in the English language. An example of the model's performance can be seen in the table below.

Comment	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
If I could make social media disappear, Tik Tok would be the first to go.	0.016813	0.005273	0.009925	0.004632	0.944276	0.003072	0.016008
Why is it propaganda for someone to say they're voting 3rd party? Because you don't like their choice? This country's duopoly party system is awful and I refuse to participate in it any longer. I'd sooner vote for RFK jr over any RNC or DNC nominee	0.341550	0.387735	0.012678	0.016066	0.227679	0.010913	0.003380

Table 16

As can be seen in these examples, neutrality is a big part of the sentiment that comprises a comment. In the case of the first comment, it can be seen that the overall emotion present in it is neutral regardless of the fact that it speaks ill about the social media platform. While in the second comment, the model assigned a score of 0.342 and 0.388 to the anger and disgust variables respectively, however it still assigned a score of 0.228 to the neutral variable, meaning that neutrality is considered a big part of most sentences by the model.

For this part of my analysis, I figured it would be interest to assign a label to each comment according to the sentiment value that was the highest for each comment and present a count of the overall emotions present in the comments for both groups of subsets.

However, for this part I decided to do a first count while including the neutral emotion as sometimes, it might be genuinely unclear whether the comment expresses a specific emotion, and having a neutral option would allow for more nuanced categorization. I then did a second count without considering the neutral emotion to allow for those underlying emotions to come out to give us a balanced idea of the overall way people feel.

Trump subsets.

A first look at these bar plots in figure 11 reveals that most of the comments from each of the subsets that mentioned the former president by his last name convey a mostly neutral emotion in their statements, while the most prominent emotions after that are disgust and anger respectively, while sadness and joy are the least present in them.

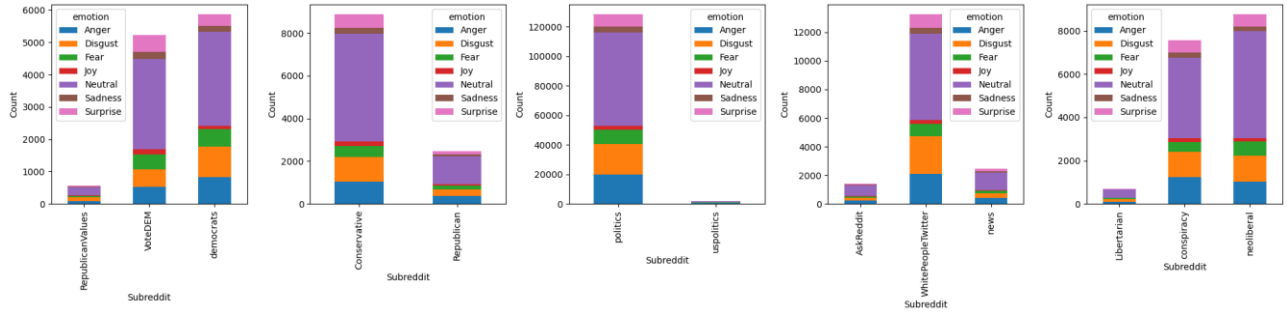

Figure 11 Emotion scores for the Trump subsets (neutral included)

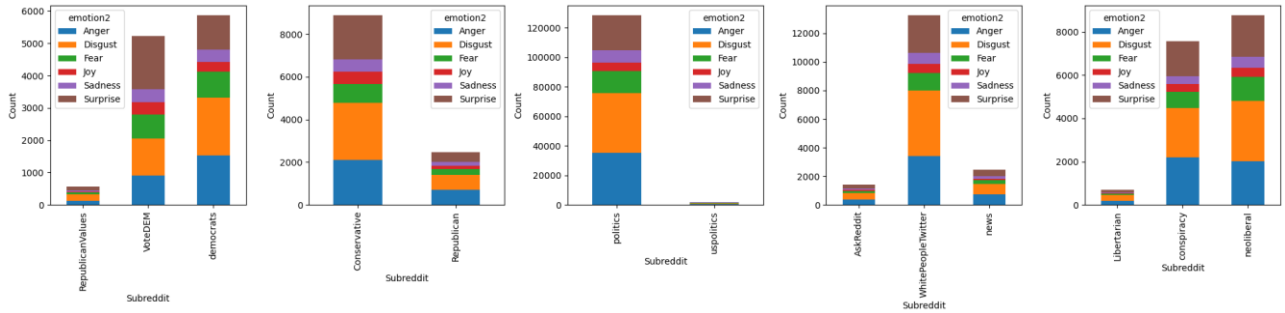

Figure 12 Emotion scores for the Trump subsets (neutral excluded)

However, taking out the neutral emotion out of the equation reveals that comment which were mostly neutral harbored feelings of disgust, anger, and surprise respectively given that all these three values effectively doubled and, in some cases, even tripled their numbers.

Sadly, it is a bit disappointing to see that all communities share more or less the same number of comments for each particular emotion, without one being more prevalent than the other. The only thing that can be noted is the fact that a small majority of comments in the *r/VoteDEM* subreddit expresses surprise as their main emotion, something that doesn't happen in the other communities.

The following are some of the most emotional comments in this group.

Comment	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	subreddit
honestly i'm here for it bc it makes up a bit for these disgusting Trump +6-8 polls in the average from Rasmussen, Harris, etc.	0.0034	0. 991	0.0018	0.0046	0.0016	0.0010	0.0004	VoteDEM

https://www.reuters.com/world/us/exclusive-fbi-finds-scant-evidence-us-capitol-attack-was-coordinated-sources-2021-08-20/ Why are you still arguing with feelings? You're definitely angry you have no rebuttals to the facts I'm providing you. Question, was Trump still talking when people started rioting at the capitol?	0.9900	0.00134	0.00181	0.00048	0.00280	0.00157	0.00195	AskReddit

Table 17

GOP subsets.

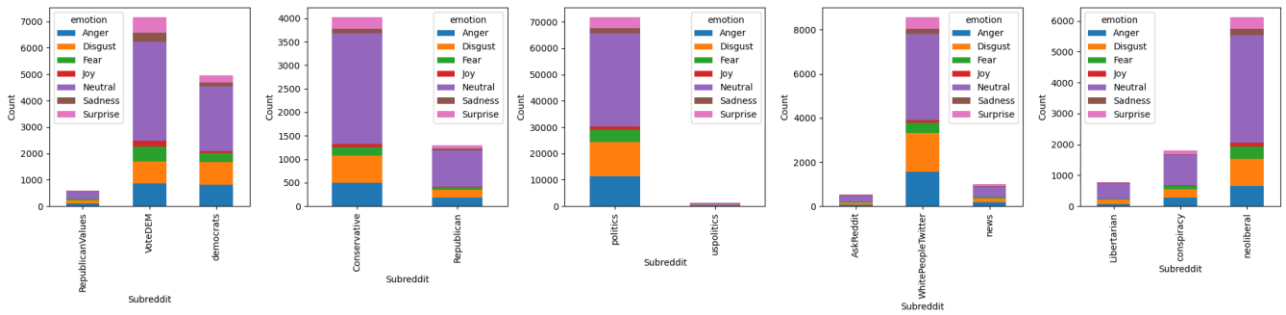

Figure 13 Emotion scores for the GOP subsets (neutral included)

Looking at the results in the GOP subset, shows similar counts for the emotions as in the Trump subsets. However, it is important to note that there is much less disgust and anger present in the democrats subset as well as in the *nutjobs* subset when compared to those values in the Trump group.

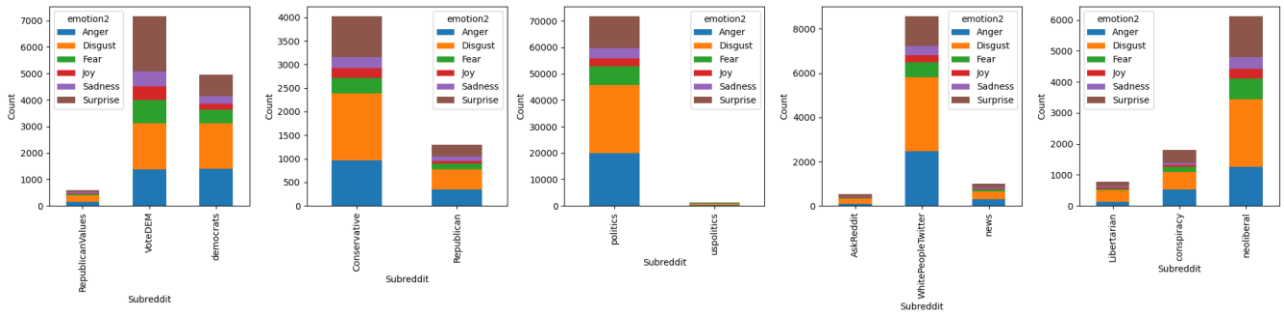

Figure 14 Emotion scores for the GOP subsets (neutral excluded)

Once again, a second pass was made to find out what the underlying emotions were. This time it revealed that there is in fact a lot of disgust, anger and surprise in the comments that mention the GOP directly, a fact that seems perhaps unsurprising given the way the discourse has revolved around the Republican Party in the last few months.

Table 18 shows some examples of highly emotional comments.

Comment	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	subreddit
Since the Republicans desire to make the U.S look weak and unreliable is working, Europe and supporting nations will have to pick up the slack #sad	0.0015	0. 0003	0.00119	0.00299	0.0013	0.99013	0.00255	democrats
I'm surprised that DeSantis hasn't realized that the Moral Majority Republican voters have largely faded and that is isnt helping him win the presidency.	0.00358	0.001265	0.00111	0.00170	0.006098	0.00155	0.98470	Conservative

Table 18

7 Conclusion and further work

After having done sentiment analysis on these comments, one thing that became apparent was the subtlety with which an opinion can be shared and how an NLP model may or may not be able to pick up on them. This fact became apparent when I first came across an unexpectedly large number of comments being classified as positive by the distilBERT and Text Blob models in the democrats subset in both the Trump and GOP groups. The fact that these two models consider non-negative comments as inherently positive can lead to misunderstandings and misinterpretations of their analyses.

The power of these subtleties became even more apparent when looking at the results of the emotion analysis done with the distilRoBERTa model, which allowed me to see the amount of mostly neutral comments there were and made me realize how nuanced a written message can be.

All in all, all these models have their applications and are more or less informative depending on the context they are used in. For instance, in the case of online product reviews, models like Text Blob and distilRoBERTa cause no problems by lacking a neutral label since a review is either positive or negative.

However, in the specific case of Reddit comments, this type of results only shows half of the truth since they are able to identify negative comments correctly while unintentionally inflating the numbers for positive comments given the fact that they consider non-negative comments as simply positive.

This is why the approach that the VADER model takes seems to me to be the most useful for analyzing the sentiment of Reddit comments, since what it does is break down the entire sentence, assign a score to the individual pieces and the calculate a compound score which gives a greater weight, and therefore, importance to both the negative and positive words and giving us a value that actually represents the sentiment in the text.

Lastly, I would like to add that, even though machine learning has made huge leaps in the variety of problems and situations, one has to be careful when working with Natural Language Processing, as the results obtained from it can widely vary and may mislead people into thinking something is different from what it actually is, and we must always remember that these models are only as good as the data they are trained on.

8 Bibliography and sitography

Zeihan, P. (2022). The End of the World Is Just the Beginning: Mapping the Collapse of Globalization. Harper Business.

Liu, Y., Ott ,M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

Hashemi-Pour, C., BERT language model, https://shorturl.at/gE578

Distilbert Community, distilbert-base-uncased, https://huggingface.co/distilbert/distilbert-base-uncased

Gong, A. AG6019/reddit-comment-sentiment, https://huggingface.co/AG6019/reddit-comment-sentiment?text=I+like+you.+I+love+you

Asaniczka, Public Opinion on Republicans (Daily Updated), https://www.kaggle.com/datasets/asaniczka/public-opinion-on-republicans-daily-updated

Problem Report Corpus, https://www.kaggle.com/datasets/nltkdata/problem-report-corpus

Open Multilingual Wordnet, https://omwn.org

Mitchell P. Marcus et al., Penn Treebank, https://paperswithcode.com/dataset/penn-treebank

NLTK Documentation, https://www.nltk.org/_modules/nltk/corpus/reader/lin.html

https://www.reddit.com/r/democrats/about/

https://www.reddit.com/r/VoteDEM/about/

https://www.reddit.com/r/RepublicanValues/about/

https://www.reddit.com/r/Republican/about/

https://www.reddit.com/r/Conservative/about/

https://www.reddit.com/r/politics/about/

https://www.reddit.com/r/uspolitics/about/

https://www.reddit.com/r/WhitePeopleTwitter/about/

https://www.reddit.com/r/news/about/

https://www.reddit.com/r/conspiracy/about/

https://www.reddit.com/r/Libertarian/about/

https://www.reddit.com/r/neoliberal/about/

https://www.reddit.com/r/neoliberal/about/