Sentiment Analysis on Twitter Classification on gun control: Democrats vs Republican

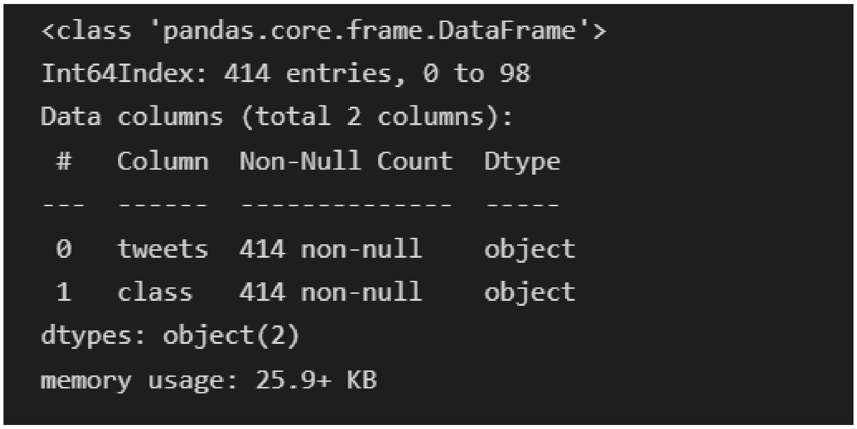
Hyeonu(Eric) Kim: 301338435
Jooeun Park: 301414492

Introduction

With the ongoing debate on gun ownership between Democratic supporters and Republican supporters, we wanted to see the ratio between each supporter on gun ownership from twitter. What types of words were most used amongst these three groups (Neutral, Democratic, Republican). Ultimately to classify what type of party they are supporting just by looking at their tweet, we will achieve this goal using the BERT model.

Data Collection

Data gathering is done by python bot implementing twitter API. The most recent 100 search results for all users are saved in a csv file, and 600 tweets are gathered in total. Duplicated tweets, retweets, mention IDs, and data that can’t be understood without conversation contexts are removed during the refinement process. 414 tweets are selected to analyze. After gathering these data, we have classified each data into N, D, R for Neutral, Democratic, Republican tweets after studying what each political stances are as shown in Fig1. Note that these could be the crucial errors on our part as we classified these could be severely subjective and could affect our results.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 414 entries, 0 to 98
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   tweets  414 non-null    object
 1   class   414 non-null    object
dtypes: object(2)
memory usage: 25.9+ KB
```

Figure 1. Attributes of our dataset.

BERT Model

For this project, we will try to replicate a set of experiments involving semantic analysis on tweets and customer reviews. People have already made an attempt, however we believe we may find new findings. Sentiment analysis is a growing area of Natural Language Processing with research from document level classification (Pang and Lee 2008) and Twitter Emoticon Analysis (Kouloumpis, Wilson, Moore, 2021). Twitter semantic analysis has been performed before (Kale, Padmadas, 2017) and research into how to preprocess tweets as well (Ramachandran, Parvathi, 2019) "BERT model is a model that learns contextual relations between words in a text" (Horev). BERT takes a sentence with tokens on [CLS] in front and [SEP] at the end to indicate the start and finish of the sentence. The BERT model also consists of Tokenization and Word Embedding.

"Tokenization helps breaking the text into tokens in which we convert the sentence from a list of strings to a list of vocabulary indices" (Horev). This allows us to calculate what the most common words that appear in sentences are to help with classification.
Word Embedding represents words/texts into a number of vectors. This allows us to boost the performance in NLP tasks.

Data Analysis

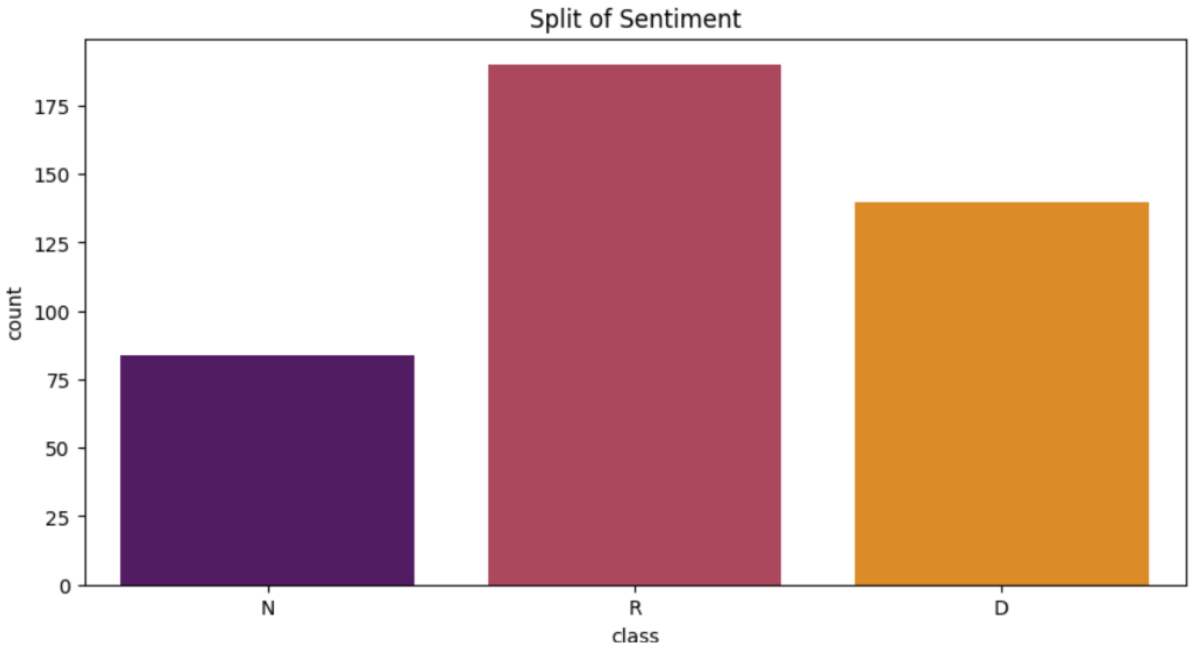


Figure 2. Ratio between each classes

As shown in Figure 2 above, there are a lot more Republican and Democratic tweets than Neutral tweets. There are also more Republican tweets than Democratic tweets regarding gun control. This could be due to the time we have gathered our data where a US government has introduced new laws regarding gun control.

In addition, the most frequent 15 words are selected and analyzed by number of appearances. For clearer results, we removed the articles and be-verbs from the word lists. As shown in the bar plots below (figure 3, 4, 5), the word “have” comes more frequently in democratic tweets. The word “should " only appears in the democratic’s ranking, there might be a correlation between them. It is noticeable that republican tweets have significantly more ‘laws' and ‘countries’. Republicans tend to tweet more about the laws and statistics between other nations where they do not allow gun ownership. For example, “The countries do not allow gun have higher rate of suicide.” is classified as Republican tweets, and it would affect to the result. low For neutral tweets, most of the words are just a preposition or pronouns.

It was interesting that hashtags such as #gunsens or #GunOwnersCare is not in any ranking. It could be caused by the kind of tweets we collected, where most of them are mentions (reply to other people's tweets), where people don't usually write hashtags in.

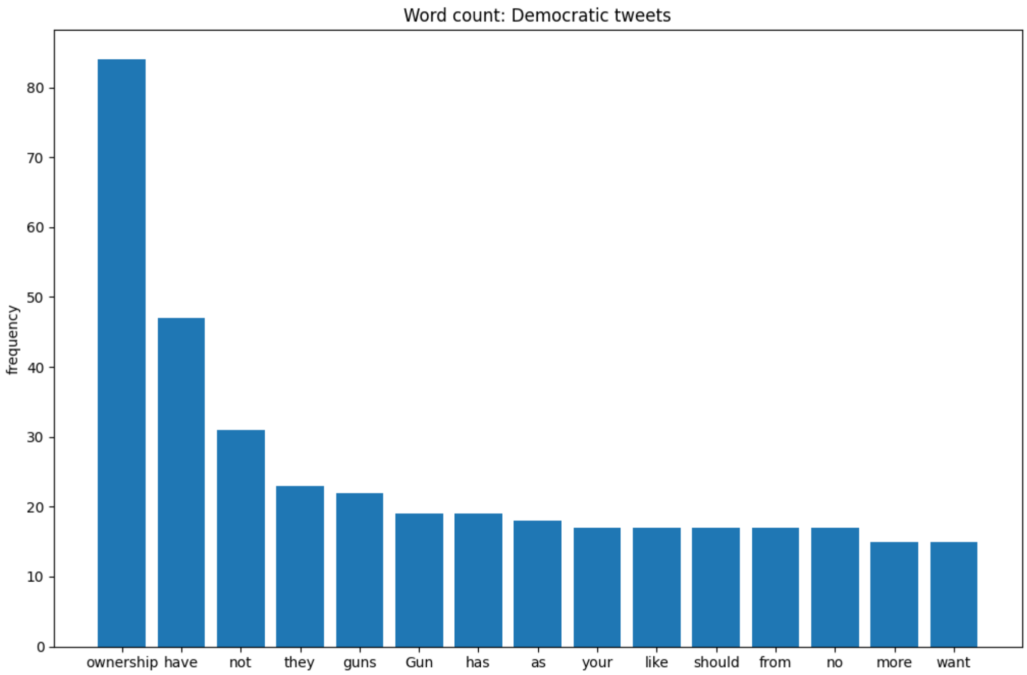


Figure 3. Frequency of top 15 words used in Democratic tweets

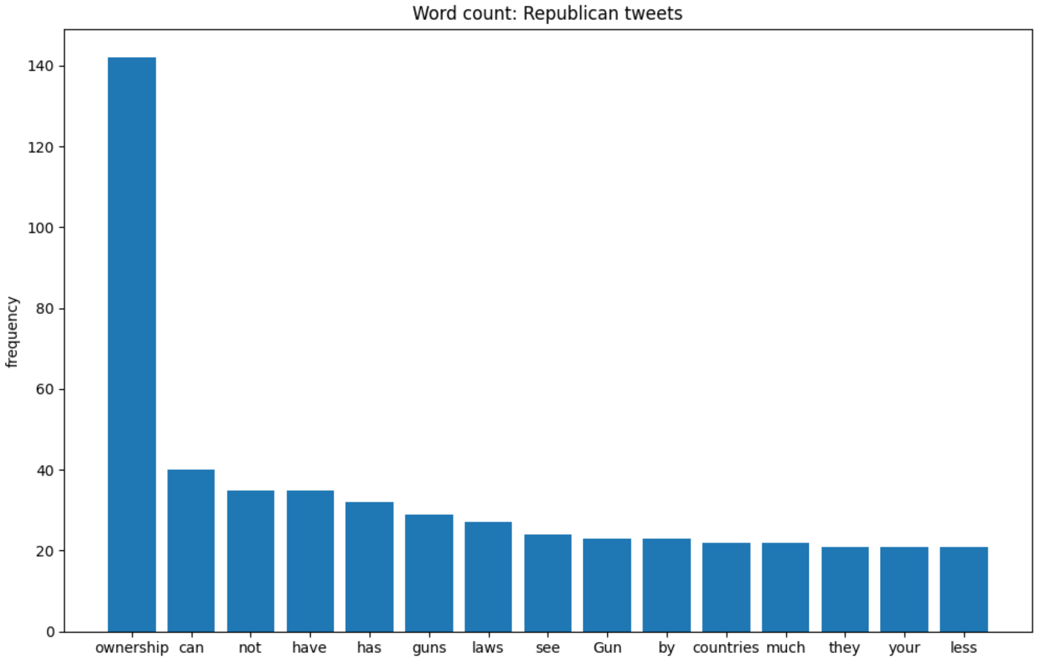


Figure 4. Frequency of top 15 words used in Republican tweets

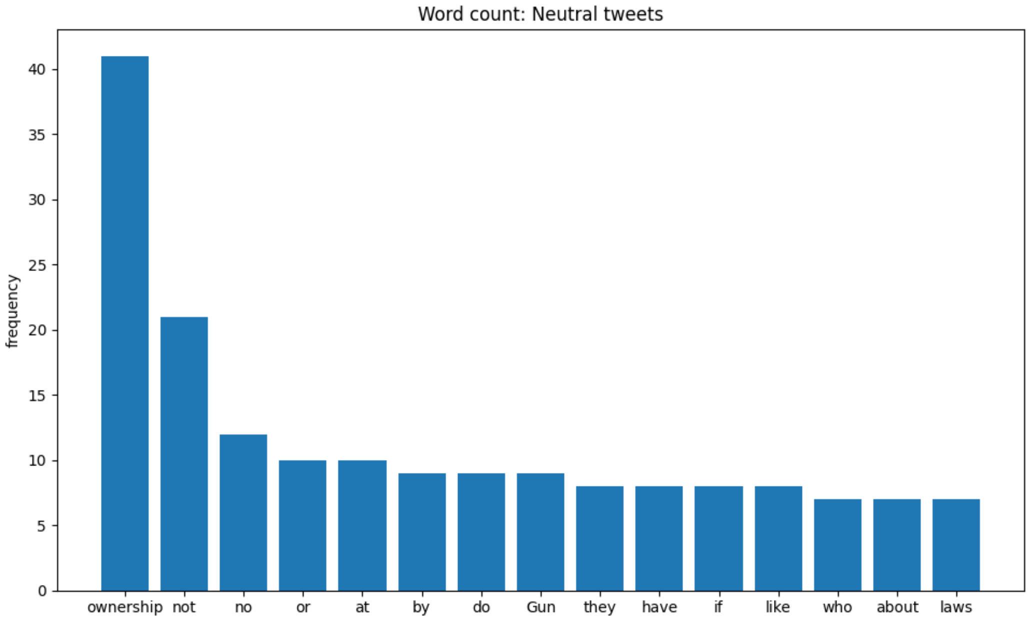


Figure 5. Frequency of top 15 words used in neutral tweets

Results

First, we used about 315 data for training and 99 data for testing. Then we had to divide our datasets from the training set for the training set and validation set using the *train_test_split* function. We then proceed to train them using the BERT model as explained in the **BERT Model** section. Below chart shows some of the experiments we did on hyperparameters and its results. We have decided to only show the changes on batch size and epochs as changing other parameters such as *MAX_LEN* for maximum length of sequence for input tokens, *lr* for learning rate of the optimizer, and different optimizers did not have much effect on results

	1	2	4	5	10
Epoch	40%	42%	43%	49%	45%

Table 1 . Chart for average accuracy on the testing set with *batch_size = 4*

	1	2	4	5	10
Epoch	48%	49%	48%	42%	38%

Table 2. Chart for average accuracy on the testing set with *batch_size = 8*

	1	2	4	5	10
Epoch	41%	40%	46%	39%	35%

Table 3. Chart for average accuracy on the testing set with *batch_size = 16*

As shown in our table, our best result is 49% accuracy on the testing set.

Secondly, we tested new sentences on the model. The model tends to give better results when the sentence includes the words that are not so common such as “parents”, “safe”, or “kill”. It gives a 40% of accuracy.

['Parents must have a gun to keep their children safe']	D
['should we have a gun ownership?']	N
['Gun ownership will kill a lot of people']	R
['I dont think we need a gun ownership']	N
['gun ownership is dangerous']	N

Table 4. Part of results

Limitations

There were only 414 tweets as a dataset, while most good models have at least thousands of dataset. It was an excessively small amount to train and verify the data. We have also classified them subjectively on how we perceive Democrats and Republicans think about gun control. As data are classified by two people, there might be contradicting labels. Those errors would have a meaningful effect on the results, due to the lack of data.

Project Experience Summary

Jooeun Park

- Classified Rebpuliacan and Democrats tweets with 49% of accuracy with the BERT model, testing various epochs and learning rates.
- Built a twitter bot that automatically collects 100 tweets each day and converts it to csv file.

Hyeonu(Eric) Kim

- Cleaned and classified data for Democrats, and Republican tweets
- Data Analysis on ratio of sample tweets we have gathered
- Researched on NLP classifications using papers and using CMPT413, 419 lecture notes and tested with different hyperparameters for BERT model

Bibliography

- Horev, R. (2018, November 17). Bert explained: State of the art language model for NLP. Medium. Retrieved December 2, 2022, from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- Bo Pang, Lillian Lee. (2008). Opinion Mining and sentiment analysis.
- Kouloumpis, Wilson, Moore. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG!
- S. Kale \& V. Padmadas (2017). Sentiment Analysis of Tweets Using Semantic Analysis