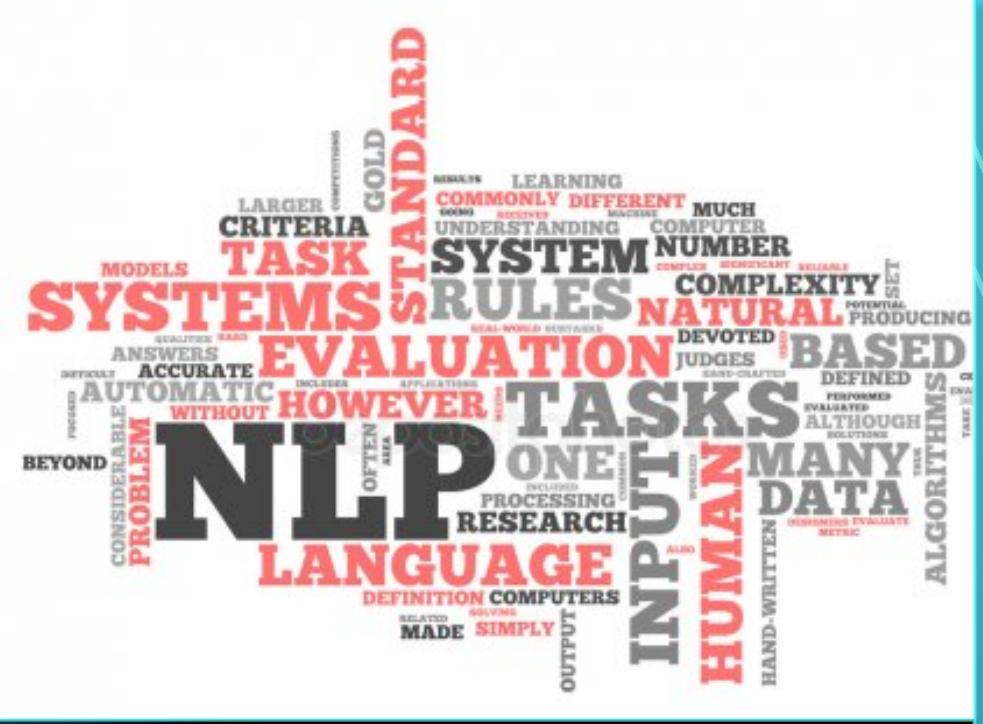

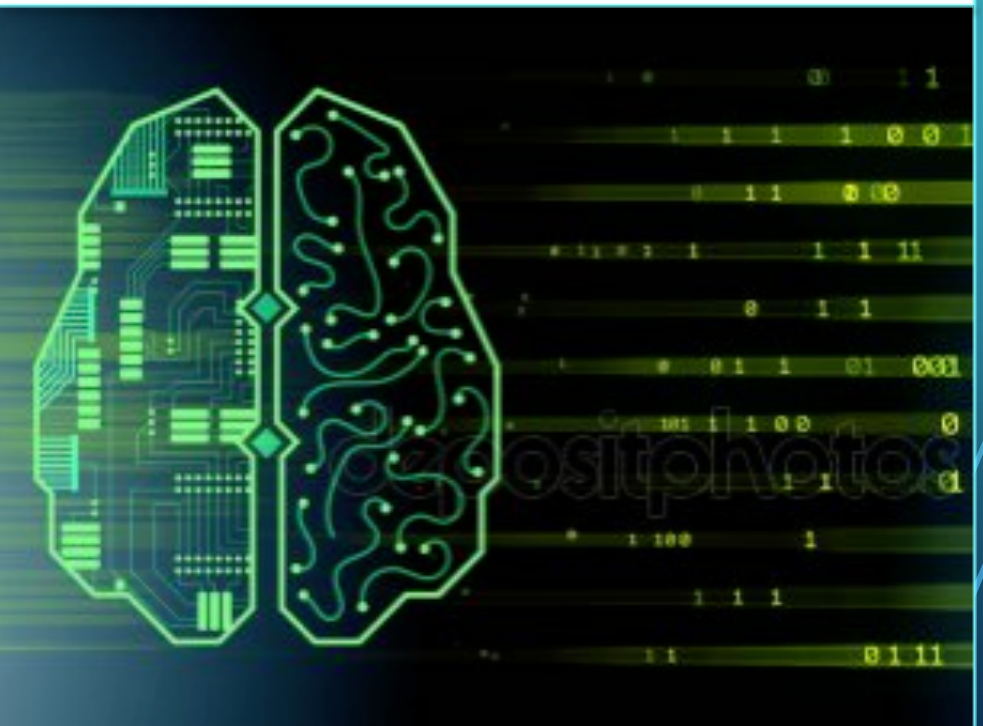

NATURAL LANGUAGE PROCESSING

COMPARING DIFFERENT MACHINE LEARNING MODELS TO PREDICT THE ORIGIN OF SUBREDDIT POSTS

Presented by
Eddie Reed
October 18, 2019

PROBLEM STATEMENT

In this project, we were asked to perform web scraping to collect subreddit posts data of two different subreddits to be used to train a machine learning model to be able to predict the origin of subreddit the posts came from. I selected the Apple and Samsung subreddits to compare. This problem can be classified as a binary classification problem. The success of the model would be determined by how accurate it performs predicting if a post came from the Apple or the Samsung subreddit by analyzing the text collected from each posts. A high accuracy score when the model processes data it has not seen before is a good indicator if the model is performing well.

DATA COLLECTION AND EDA

I used the **praw** python wrapper to interact with the reddit API to collect the subreddit posts and save them to a csv file. I collected a total of 1853 posts between the Apple and Samsung subreddits. The initial thought was to use the title and body sections of the reddit posts to train the models, however, during the EDA phase it was quickly discovered that in a large number of cases, the body section of a post contained images, links to other websites and when that data was brought into a dataframe, there would be large sections of blank space and no text data. Doing an initial pass of the data through a model yielded that the 'body' section of the reddit submissions would not be very useful for training our model due to a large set of the data being blank.

After further evaluation of the data, it was determined that the 'title' section of the subreddits would be the best data to use because each row in the dataframe contained sufficient text data.

PREPROCESSING AND MODELING

- We were asked to evaluate at least two machine learning models to process the text data we collected and determine which one performs the best in predicting the origin of a subreddit post. We were required to use at least one model from the Naive Bayes model family and we could select any other classification model of our choosing (e.g. logistical regression, Support Vector Machines (SVM), KNN, etc.).
- Since we are analyzing text data, we had to use Natural Language Processing (NLP) techniques to convert the data into some vectorized format our classifications could understand and process. We learned about two NLP transformers called CounterVectorize and Term Frequency-Inverse Document Frequency(TFIDF) vectorizer. Both models will analyze text data and transform it into sparse vectors that our classification estimators can process.
- As in all cases in preparing to model, once cleaning and EDA of the data is completed, it is now ready to be fit into our model. We split our data into a training and data set using train_test_split, fit and score our model on the testing data. We have several hyperparameters we can tune based on the model we are using to validate and improve our models.

EVALUATION OF THE MODELS

In this project, the following machine learning classification models were used:

Logistic Regression	Bernoulli	Multinomial NB	Support Vector Machines

The two transformers used were:

CountVectorizer	TFIDFVectorizer

Results of Apple and Samsung NPL Reddit Predictions

Post Number	Actual	Predicted	Subission_Title
274	samsung	samsung	Fixing small scratch?
1118	samsung	samsung	Wireless headphones Advice
1521	apple	apple	Wall Street is underestimating how much money Apple will make off 5G, says Jefferies analysts; rates stock as buy.
258	samsung	samsung	Samsung members app broken
1464	apple	samsung	Just posting my review here as well.
188	samsung	samsung	Samsung s10 homebar
385	samsung	samsung	The Samsung Health SDK may be ruining Samsung for me
1360	apple	apple	Arstechnica iPhone 11 review
1448	apple	apple	TIL that apple keyboard can put personal info depending on a context
952	samsung	samsung	Broken power button
390	samsung	samsung	Should i get A50 or A70
809	samsung	samsung	Restore Photos To Phone?
1061	samsung	samsung	S10 overheating
389	samsung	samsung	New battery in Gear s3
1238	samsung	samsung	I'm getting the Samsung galaxy note 10+ on Thursday
1093	samsung	samsung	I joined the family! :D
1849	apple	apple	Apple Fifth Ave. Store Reopening Animation
375	samsung	samsung	Do you think Samsung will release a compact S11 such as the S10e?
769	samsung	samsung	Hi everyone, I just bought the note 10 plus, and I realized that it's not showing true blacks in the display. Just really dark greys, and I cant find anything online that would suggest this isnt usual.
129	apple	samsung	Honestly I think Apples repair center is a lawsuit waiting to happen
1270	samsung	samsung	Top left of s9 back cracked, water resistance compromised?
587	samsung	samsung	Samsung a70 problem
892	samsung	samsung	Screen completely off

KEY TAKEAWAYS

- Bernoulli model was 95% accurate in predicting reddit posts.
- Only 24 posts we misclassified out of 464 posts from the test set.
- Smaller dataset was used to train model which saved time and money.

QUESTIONS?