Udacity Machine Learning Nanodegree 2019

Capstone Proposal:

Understanding Emotion in 280 Characters

Daniel Blignaut

June 2019

Domain Background

Since the advent and explosion of social media as a form of communication, an entire new touch point and data set has been created, one that is bigger in size than others and generates an astonishing amount of content daily. So much so in fact that over 500 million tweets and potential data points are created every day (Cooper, 2019).

Understanding this pool of data and using it to make inferences and commentary on what is happening in the world is a powerful tool. This realm of understanding, known as computational linguistics in the 1990's and natural language processing in todays terms has skyrocketed into a key field in artificial intelligence. The main reason for this has been a series of major breakthroughs but even more importantly is the large increase in text available online for processing. A statistic describing this is the fact that over 99% of research papers on this topic were published after 20014 (Mika V. Mäntylä)

Sentiment analysis, a particular subsection of natural language processing attempts to distill emotion and attitude out of a body of text using various methods. The applications of this are endless. Sentiment analysis is used in fields such as:

- **Customer satisfaction**: tracking customer happiness over the user's lifecycle. As businesses scale and the number of customers increase, it becomes difficult to manually interpret the satisfaction of users (Reptation.com editors, 2018)
- **Customer service prioritization**: By analyzing inbound requests, ticketing systems can make decisions on which tickets to prioritise based on the customers emotional discontent (Reptation.com editors, 2018)
- **Politics**: Using social media information to understand public opinion on specific political groups and recent decisions (Bermingham)

In more recent years, sophisticated solutions to sentiment analysis have been developed. A general approach for sentiment analysis is the following:

- **Data preprocessing**: Natural language is an extremely verbose and nuanced set of data. For example, there are many variations of expressing the same term. Due to this, we need to first preprocess and "clean" our data. In NLP, techniques such as bag of words is used to vectorise sentences and other algorithms such as an implementation of Zipf's law are used. Zipf's law explains that the frequency of any word is inversely proportional to its rank in the frequency table. For example, common words such as, "a" and "the" are less important than words like "love" which express emotion in our case. Therefore it's important that we convert this in our dataset. (Wikipedia, n.d.) (Shah, 2017)
- **Machine Learning implementation**: Traditional methods then employed Naives Bayes, maximum entropy or SVM methods. Depending on which method chosen may impact our data preprocessing step. For example, if we choose SVM, we may use a unigram feature extractor and turn a tweet or sentence into two sets of vectors, the first containing each relevant word and the second containing a binary 1 if the sentence contains relevant word and a 0 if it doesn't. SVM would take these vectors as an input, apply a specified kernel such as a linear kernel and output a classification. (Alec Go)

My personal motivation for this project goes above and beyond this capstone. I intend to use the findings of this project for subsequent research. I have a background in accounting, finance and software engineering. Due to my thorough understanding of financial instruments and the underlying finance, all of these products are based on no-arbitrage and efficient market hypothesis conditions. It is my personal belief that these products, particular items such as exchange rates who's value is not based on long term intrinsic company value, is much more likely to be influenced through behavioral finance and sentiment. I plan to use my findings from this project to explore the cause effect relationship between FOREX and tweets on certain topics and whether or not FOREX prices are subject to influence from tweets or whether tweets are subject to influence of the economic climate.

Problem Statement

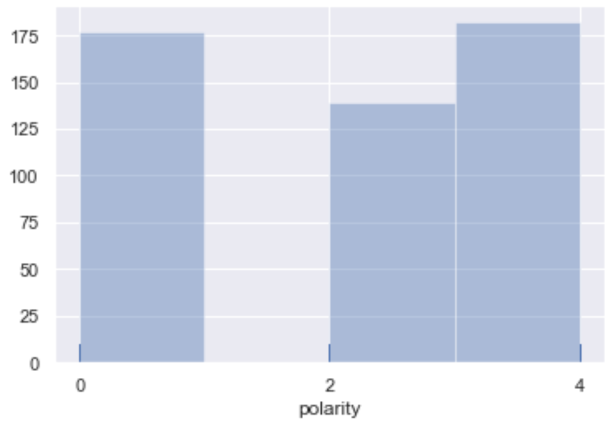
The objective of my project is to analyze a dataset of tweets from twitter and to classify them as positive, negative or neutral sentiment. When given a tweet from twitter, it should first clean the data up and turn it into an appropriate input for the model. The model itself should then output either a probability for each category or a single value that is within a certain threshold. This should then be interpreted as a final sentiment.

Datasets and inputs

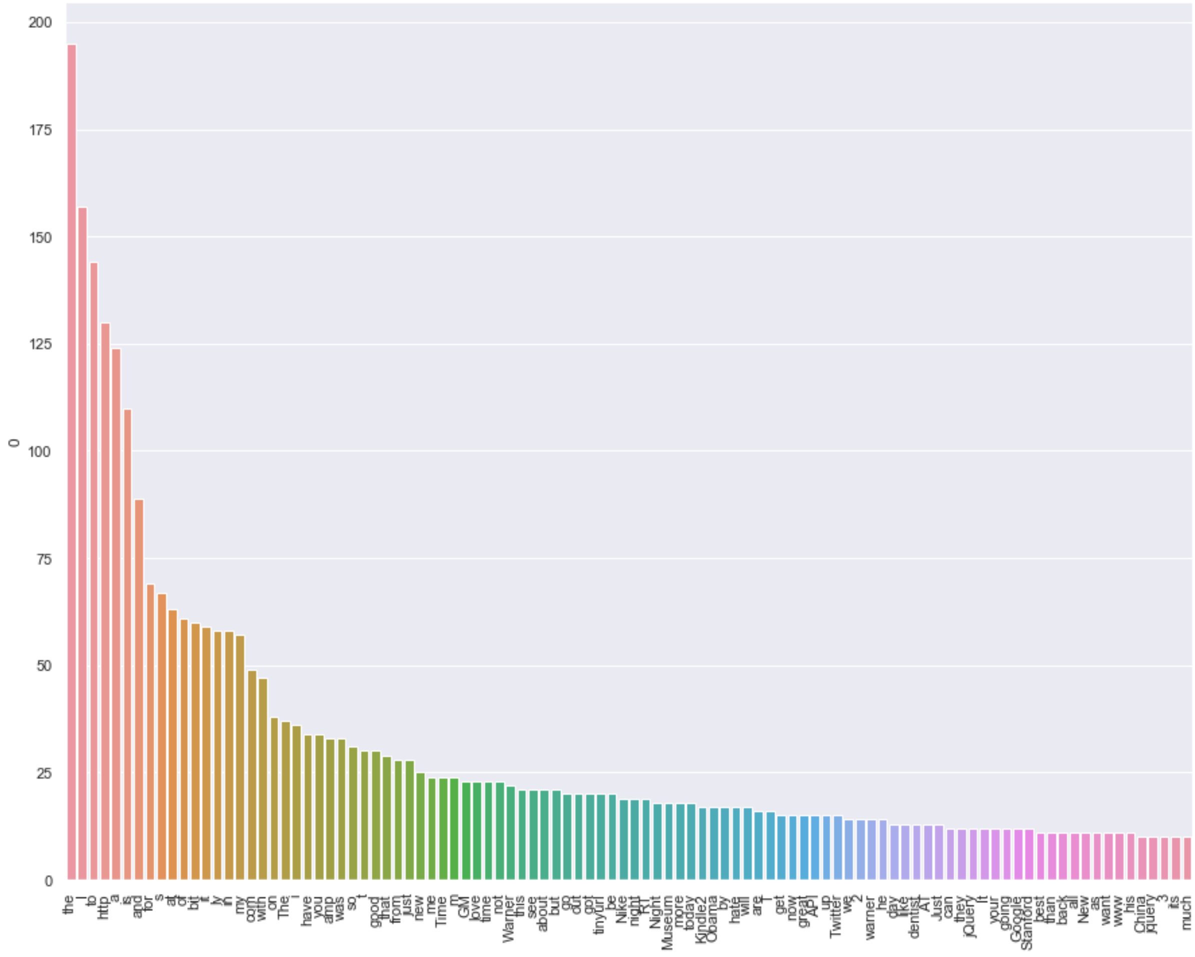
For this project, I chose to use the Sentiment140 dataset. This is a dataset that originates from Stanford University and contains over 1.6 million data entries. Each entry has the following columns for the corresponding tweet: (University, n.d.)

- Sentiment
    - 0 – negative
    - 2 – neutral
    - 4 – positive
- Id
- Date
- The query (lyx). If there is no query then the value is NO_QUERY
- The user that sent the tweets (their handle)
- The text

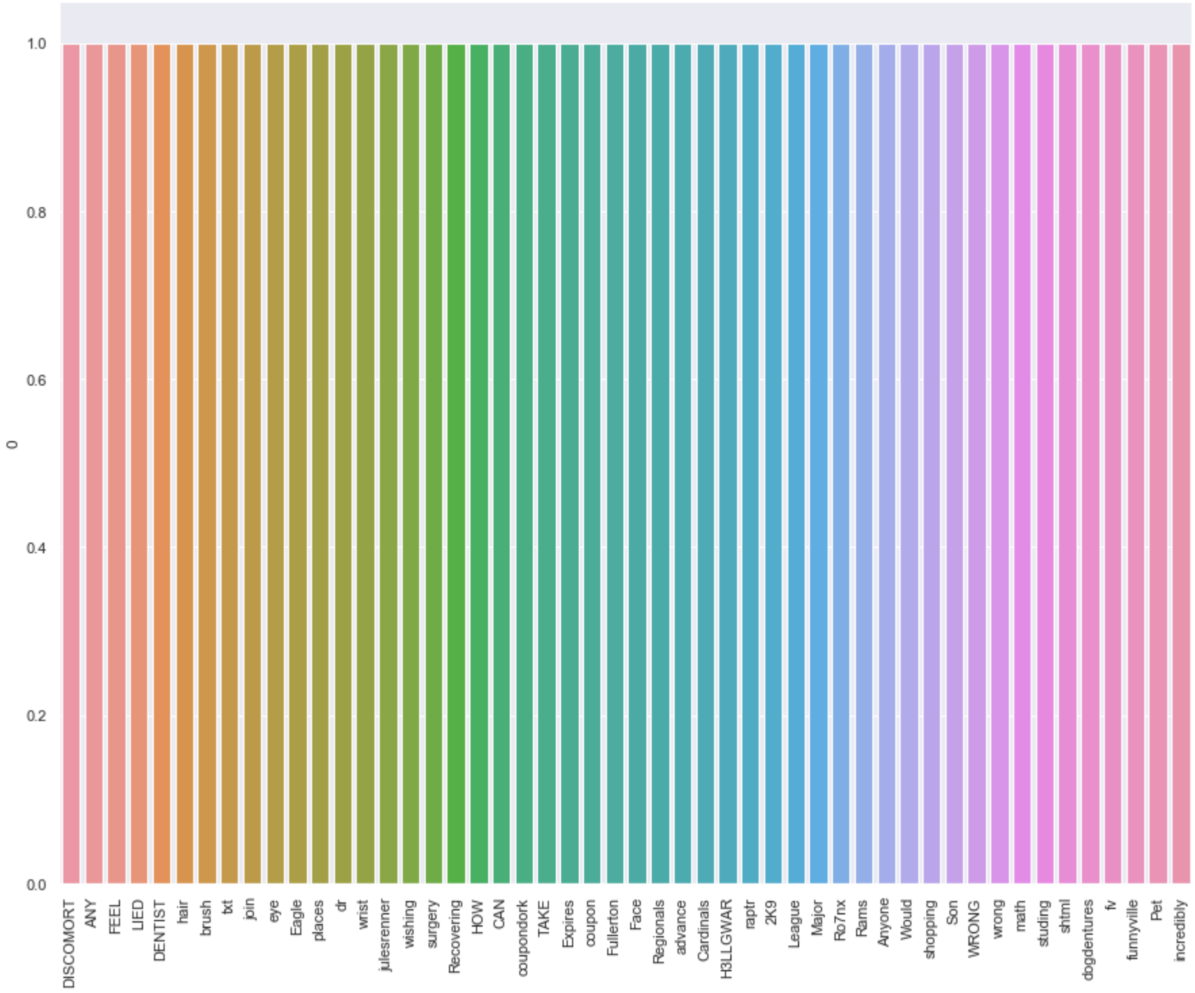
Test data set polarity distribution



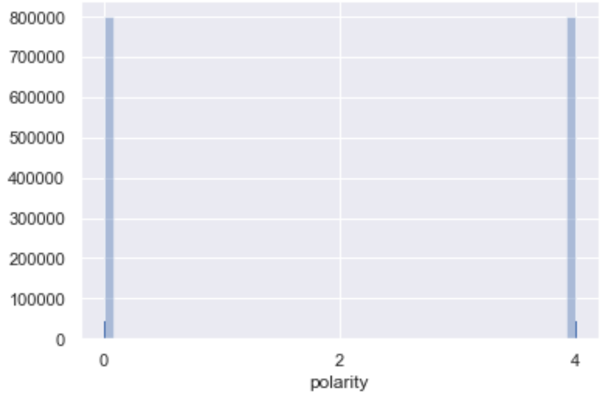
Test data set most used 100 words



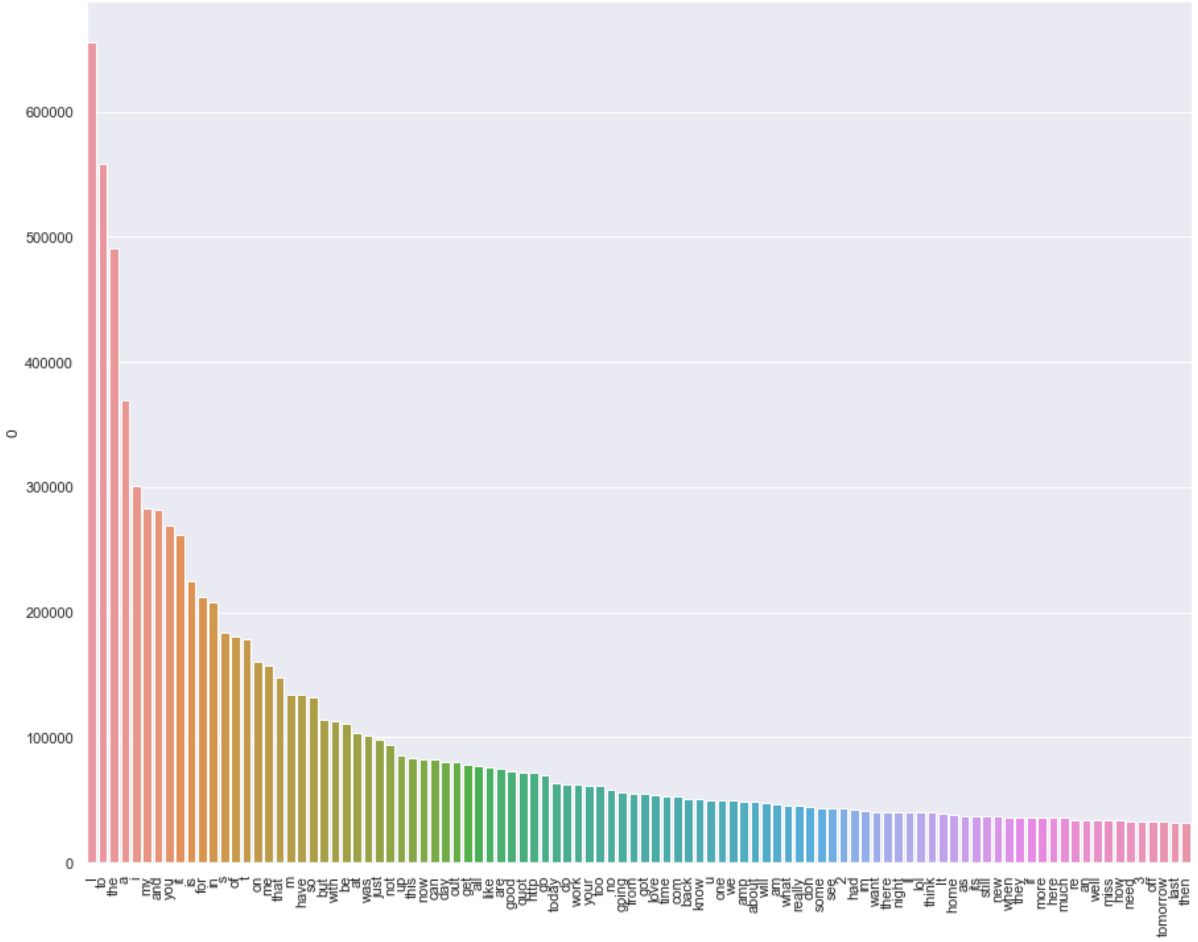
Test data set least used 100 words



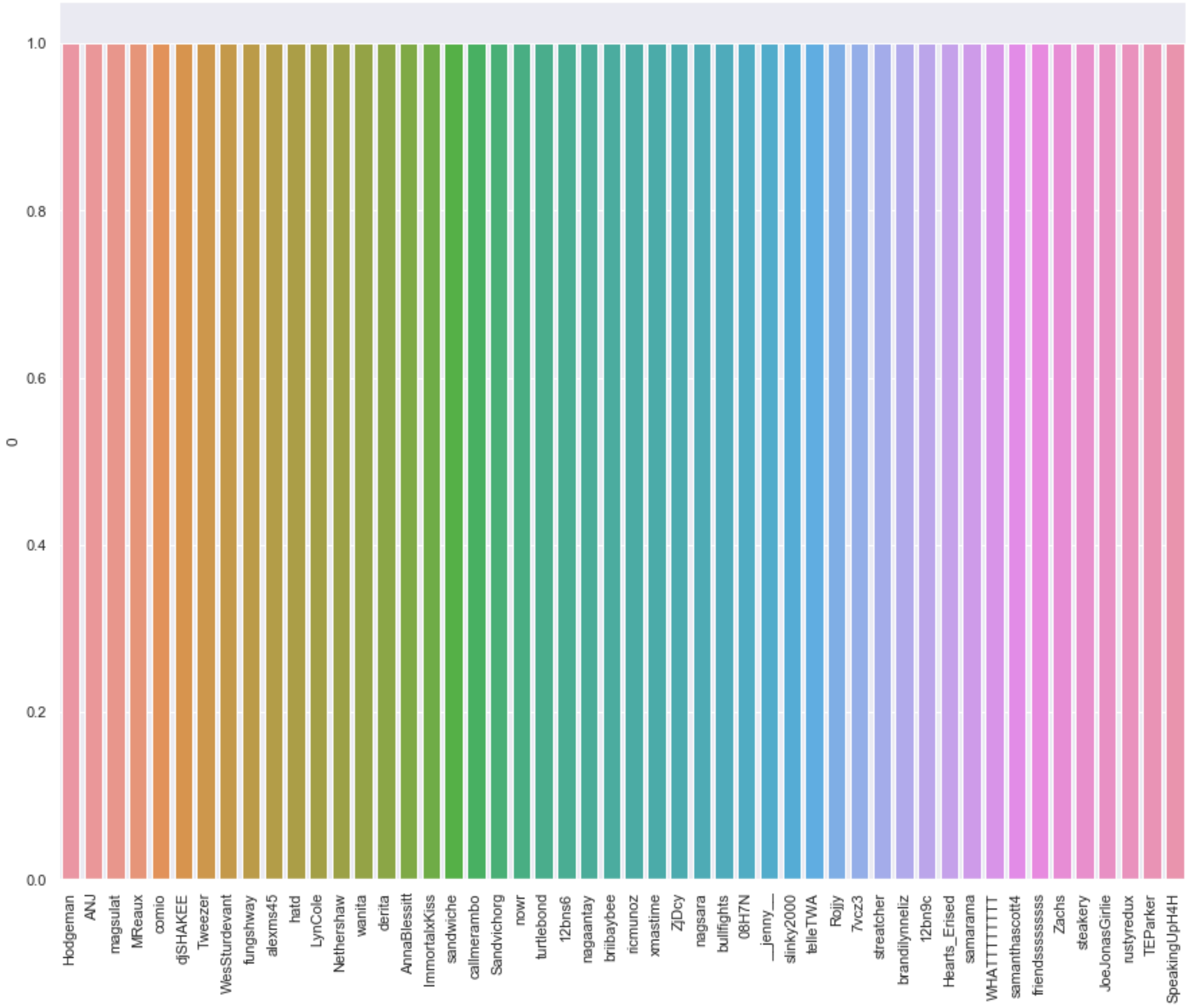
Training data set polarity distribution



Training data set most used 100 words



Training set least used 100 words



As can be seen from above, the training data set is evenly distributed between positive and negative rated answers. The test dataset however contains a third category, neutral which contains roughly 40 less data points than the positive and negative labels (which have the same distribution).

Upon further analysis, the importance of data prepossing is extremely relevant. It is clear in both the training and test datasets that the most highly distributed words are those which occur most often (by definition) but also imply the least in terms of sentiment (I, the, to, and my). Another issue is that in the elast most common set of words, user handles occur frequently, these also don’t indicate sentiment and should be removed.

Solution Statement

I intend to use deep learning techniques to try and solve this problem which have been used in recent years with success to improve sentiment analysis. More specifically, I’d like to implement a Long Short Term Model which is a form of Recurrent neural networks with the unique ability to learn long-term dependencies. I believe an implementation of this model may be useful in helping to better formulate a solution that understands the sequence of past data it has interpreted and can use this when applying sentiment to a word or word vector.

However, due to the subtlety of the problem and the skill required to classify sentiment correctly, I'd also like to focus on a few other key components of the project, including word vectorization and the various techniques we can possibly use here to test the model.

Benchmark Model

As a benchmark model, I plan to use an SVM supervised learning model with the same data pre-processing discussed above. SVM, Naives Bayes and Max Entropy have shown historical success in the problem with scores in the 60-70% range for accuracy. (Alec Go)

I will include an implementation of this model in the report with the corresponding scores.

Evaluation Metrics

The main metric I plan to use to evaluate these models is f1 score. However, I also intend to closely analyze the following:

- Confusion matrix (precision & recall)
- Accuracy

Due to the relatively even distribution across categories, accuracy could be my key metric however in general I believe f1 score to be a better measure due to its relative weighting of accuracy and recall and will thus use this.

Project Design

I intend to break my project down into the following sections:

Data pre-processing

This section will include exploration on topics for pre-processing the dataset, including:

- Conversion to lowercase
- Normalizing query terms
- Removal of handles
- Removal of URLs
- Removal of overly repeated characters
- Removal of numbers
- Expansion of apostrophe words (e.g. can't = cannot)
- Remove stop words
- Remove emoticons

Word vectorization

This section will cover various word vectorization techniques commonly used in sentiment analysis, including:

- Unigram
- Bigram
- GloVes

It will also cover visualizing some of these vectors and their statistical inferences to get an understanding on any further pre-processing necessary before feeding them into the respective models. Finally, we will create cross-validation sets of our data for training.

Model training & testing

In this section we will create implementations of:

1. SVM benchmark model
2. LSTM recurrent network model

We will iterate through various hyperparameters for these models by making use of our cross validation datasets to try and maximize the accuracy score.

Data visualization

We will then focus on visualizing the metrics we are focusing on.

Bibliography

Cooper, P. (2019, 01 16). *Twitter Statistics*. Retrieved from Hootsuite: https://blog.hootsuite.com/twitter-statistics/

Mika V. Mäntylä, D. G. (n.d.). *The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers.* Retrieved from arXiv.org: https://arxiv.org/pdf/1612.01556.pdf

Reptation.com editors. (2018, December 21). *5 Real-World Sentiment Analysis Use Cases* . Retrieved from reputation.com: https://www.reputation.com/resources/blog/5-real-world-sentiment-analysis-use-cases/

Bermingham, D. A. (n.d.). *Insight.* Retrieved from Sentiment Analysis Use Cases: https://www.isin.ie/assets/38/903B8F83-111B-4DCA-A84606B4F6E557B6_document/0915_Sentiment_Analysis_Dr._Adam_Bermingham.pdf

Wikipedia. (n.d.). *Zipf's law*. Retrieved from Wikipedia: https://simple.wikipedia.org/wiki/Zipf%27s_law

Shah, D. (2017, September 15). *Exploration of one of the most enigmatic mathematical law through lens of data science* . Retrieved from Medium.com: https://medium.com/@devalshah1619/a-mysterious-law-so-simple-and-yet-so-universal-aa9f1c8903d1

Alec Go, R. B. (n.d.). *Twitter Sentiment Classification using Distant Supervision.* Stanford University.

University, S. (n.d.). *For Academics* . Retrieved from sentiment140.com: http://help.sentiment140.com/for-students/