**Final Assignment**

*Sonny Desai*

March 2020

**Identify the problem, describe data, clarify**

Journalism, especially sports journalism has become increasingly more polarized, especially on social media sites like twitter, as independent journalist vie against multi-national corporations for views, likes, and retweets. One of the main ways to move the needle, and draw attention, is to make outlandish or emotionally charged statements. In the quest for eyeballs, journalists like Grant Cohn, use their twitter account to build their brand by using outlandish language to cover their respective sports and sport teams. In assessing his particular tweets, this project seeks to use support vector machine, logistic regression, naïve bayes, and vader sentiment analysis to view the positive/negative sentiment values of each tweet. Further, the project will create an ontology and knowledge graph breaking out clusters of topics and their classification based upon tfidf scores. The data is comprised totally of tweets by journalist Grant Cohn under the user name @grantcohn. The objective is to view the sentiment scores of the tweets to see which words have a negative or positive polarity, and which tweets end up with the most favorites or retweets based upon the positive or negative sentiment score.

**Summarize algorithm/key features of code**

This assignment used a number of different models all based upon the last 3,200 tweets by Grant Cohn. To pre-process the data, I added stopwords to ourtdfidfconverter model, which converted the collection of raw tweet to a matrix of TF-IDF features, and removed the stopwords imported from the NLTK library. I also added the words “http, co, rt”, to clean up the data and remove words associated with external websites and not useful for our sentiment analyses. I also used a Lemmatizer from the NLTK library which groups together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. I then created classifiers to allow the data to be segmented into classes and to apply coefficients from our various models. I imported the training data to create our training models, here I used sentiment tokens of movie reviews and passed them through our models to get coefficient results. I created three models for this test: support vector machine, logistical regression, and naïve bayes. These models, along with the sentiment vader model, were chosen as good examples of transitioning from statistical models to knowledge-based semantically rich models. All four are used in this project to turn the text from the tweets into a coefficient from 0 to 1, based upon the polarity of the words in the tweets with positive being 1, and negative being 0. The main process in doing this is to convert the words into tokens, or singular words which made up the dataset. This was then run against the polarity models to understand their sentiment. I ran the training model, and then loaded in Grant Cohn’s tweets as this project’s corpus, and ran the aforementioned models on the dataset. Lastly, in order to help create the ontology, I create a named entity recognition model which broke down the tokens into entities, which was used to view the most common names and classes used.

**Review results**

To help view the entities that were created in the NER and svm, lr, and nb models, I created an ontology to view the most common tokens, and those that had high polarity. With Grant Cohn’s tweets as the original entity, I broke down the other classes into the most relevant entities which are quarterback, other 49ers twitter journalists, twitter users, teams, coaches, and other players. The most common entity was definitely quarterbacks followed by other 49ers twitter journalists. The highest polarity came on the quarterbacks Deshaun Watson and Jimmy Garoppolo, with Grant making comments about their good and bad play respectively. The other 49er twitter jorunalists had a lot of negative scores, but this is mostly coming from friendly banter that Grant engages with them and other twitter users. The sentiment scores have not taken into account sarcasm as of yet!

A picture containing map

Description automatically generated