Capstone 2 Proposal: Valery Lynn

Title: National patterns in race tweets

Problem: Racial bias is difficult to measure due to its sometimes ambiguous and subtle presentation. Most empirical measures of racial bias or discrimination are conducted using individual self-reporting or experimental testing that attempts to obtain results using indirect or subconscious actions. A research study being conducted at UCSF, Department of Epidemiology & Biostatistics, is using Twitter data to create place-level measurement of sentiment towards racial/ethnic minorities. The study analyzes more than one million U.S. tweets related to race/ethnicity for sentiment. A sample of the tweets (6,600) have been human-labeled with 90% agreement between the researchers. The researchers have made the data available to me to help them classify the tweets as positive, negative, or neutral. A previous attempt was made, but they were unable to classify the tweets in more than two categories. They got around this by collapsing the neutral category into the negative category. This project will be my attempt to use more up-to-date NLP methods to classify the tweets into the three categories originally proposed by the researchers.

Client: The principal investigator for this study is Thu Nguyen, a post-doctoral fellow (Associate Specialist) at UCSF, Department of Epidemiology & Biostatistics. She has asked me to collaborate on this project as a data scientist. I will be given acknowledgements in the published document. I will not be compensated monetarily for the work. The background information for this proposal was obtained from an unpublished draft of the study results given to me by the author.

Data: The tweets were obtained from March 2015– April 2016 using Twitter's Streaming Application Programming Interface (API) to continuously collect a random sample of publicly available tweets. This API gives access to a random 1% sample of tweets. The data was restricted to only tweets with latitude and longitude coordinates for the contiguous United States and the District of Columbia. In total, there are 79,848,992 million general topic tweets from 603,363 unique Twitter users in the dataset. A keyword list of 398 racial/ethnic group terms and slurs was derived from racial and ethnic categories used by the U.S. Census and the online database of racial slurs. These keywords were used to identify tweets that contained at least one race-related word. This resulted in 1.25 million tweets. Out of this, the training set was created by randomly selecting a subset of 6,600 tweets for manual labeling. Three coauthors labeled three possible responses: negative, neutral, or positive. Inter-rater reliability

among the coders was 90% for sentiment rating. In addition, the labelers indicated whether the tweet used discriminatory or stereotyping language about a racial group, whether the tweet mentioned a non-food location (e.g., Jewish Center), and whether the tweet should be excluded because it was not race-related (relating to individuals, people, or culture).

Approach: The data has been preprocessed for analysis. Python and an R-tree were used to build a spatial index to create a spatial join of the tweets. Each tweet was divided into tokens using the Stanford Tokenizer18, an open access software tool that divides text into tokens, which roughly correspond to “words.” Previous sentiment analysis was conducted using the Stochastic Gradient Descent Classifier in Python software version 2.7. This is a linear classifier that has the value “0” or “1” and is appropriate for classifying into only two groups. To accommodate these restrictions, the researchers joined the neutral category with the negative category as this gave more accurate results in preliminary analyses.

Tweets have been preprocessed to remove any stop words (e.g., the, a, is), urls, Twitter username references, additional white spaces. All words have been converted to lower case. Two or more repetitions of a character were replaced with the character itself. More than two repetitions of a word were replaced with two repetitions of that word. All punctuations and hashtag symbols were removed. In addition, a count vector was constructed for unigrams (1 word) and bigrams (2-word sequence) in the preprocessed tweets. TFIDF values were calculated for each of the words in the training dataset.

I will attempt to create a classifier for three categories using Sentiment Analysis, which “refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (Sentiment Analysis, 2018).” One algorithm that is performing well for this is the Multinomial Naive Bayes (Bag of Words) Model. This is a probabilistic model based on the labels in the training data (i.e., it classifies based on the probability of a certain label given the feature data). Probabilistic models allow for a nonlinear decision boundary and therefore more than two categories are possible. Two other classification algorithms will also be tested, Boosted Trees (gradient boosted decision trees) and Random Forests. These also have more fluid decision boundaries and have proven well for classification. In fact, a comparison of classification algorithms showed that Bayes classification was outperformed by Boosted Trees and Random Forests (Caruana & Niculescu-Mizil, 2006). However, Naive Bayes models are proving themselves to work very well in sentiment analysis, and in particular are being used successfully to analyze Tweets (Smetanin, 2018).

Deliverables: Code: data visualization, exploratory data analysis, machine learning algorithms, predicted and classified Tweets. Report on the project. Presentation.

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. doi:10.1145/1143844.1143865

Sentiment analysis. (2018, December 10). Retrieved from https://en.wikipedia.org/wiki/Sentiment_analysis

Smetanin, S. (2018, September 01). Sentiment Analysis of Tweets using Multinomial Naive Bayes. Retrieved from https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b