A close up of a sign

Description automatically generated

ALY 6080: Integrated Experiential Learning

Individual Project Proposal: 2nd Draft

Submitted to

Bryce Allen

Submitted by

Md Tajrianul Islam

Date: Nov 07, 2020

*I. Abstract*

Examination of open information from web-based articulations and sentiments could yield captivating results and pieces of information into the universe of famous emotions about anything, organization or personality. The impact of Web 2.0 has incited extended activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. Accordingly, there has been an unexpected increment of excitement for people to mine these colossal resources of data for assumptions. Conclusion examination or Opinion Mining will be mining of feeling polarities from online web-based media. In this undertaking we will discuss a methodology which licenses use and comprehension of twitter data for assessment examination. We play out a few stages of text pre-handling, and afterward try different things with various order systems. Utilizing a dataset of 1.6 million tweets and TFIDF highlights, we examination the precision acquired utilizing different classifiers for this undertaking. We locate that direct SVMs give us the best precision results among the different classifiers attempted. Supposition examination classifier could be helpful for some, applications like market investigation of various highlights of another item or popular assessment for another film or discourse by a political up-and-comer.

*II. Introduction*

Our task is engaged after breaking down the huge dataset of tweets to have the option to characterize them as sure, negative or nonpartisan, known as Sentiment Analysis. It's otherwise called assessment mining, alluding to the utilization of characteristic language preparing to decide the disposition, conclusions and feelings of an essayist or other subject inside an online notice. One of the most established technique utilized in enormous information investigation is the Natural Language Processing otherwise called NLP, is the innovation gives PCs the ability of understanding human language. Some may contend that they generally had this ability, as machine language guidance or elevated level programming dialects may meet the measures of language, yet their jargon is exceptionally little and have a profoundly organized show. What's more, it would just arrange and run where blunders are nonexistent, especially far-fetched for human dialects otherwise called common dialects which comprise huge jargon, various accents and flighty utilization of words to communicate feelings. A definitive goal of NLP is to peruse, decode, comprehend, and sort out the human dialects in a way that is significant (Garbade, 2018). The etymological wonder of refutation has been appeared to assume a critical function in Sentiment Analysis. Councill et al. [2010] tried an assumption classifier and found that including their refutation classifier gave a 29.5 % improvement in F1 score while arranging positive slant, and a 11.4 % improvement when grouping negative feeling. Kiritchenko et al. (2014) remembered a modern answer for dealing with discredited terms for their SemEval-2014 section by making tweet-explicit supposition lexica containing singular scores for terms in confirmed and nullified settings, yet the cutting edge frameworks in TSA actually utilize a basic answer for distinguishing which terms are refuted, by checking as invalidated all words from a refutation signal term to the following accentuation image. So we will be additionally concentrating how to utilize characterization to accomplish better exactness to anticipate the estimation behind a tweet.

*III. Analysis*

Our dataset for the project consists of 1.6 million tweets in english coming from two sources : Kaggle and Sentiment140. Our objective is to predict the sentimental polarity(positive / negative) for the given tweet. The basic NLP and classification techniques and tools needed for twitter sentiment analysis are explained below

**Decision Tree**

If we are given collection of labeled examples |x,f(x)| where x represents features and f(x) represents class column/label column we need to get a tree or a function that hypothesis function that approximates f(h) is a decision tree. Applications of decision trees include predicting tumor cells as benign or malignant, classifying credit card transactions as legitimate or fraudulent, classifying secondary structures of protein as alpha-helix,beta-sheet.

**K Nearest Neighbors**

This algorithm assigns a class to a new data point based on its neighbors(mode). It also identifies a numeric value of a new data point based on its neighbors(mean/median). This is also called instance based learning, case based learning, lazy learning.

Process of KNN:

* Pick a number of neighbors you want to use for classification or regression(K)
* Choose a method to measure distances (same consideration as clustering)
* Keep a data set with record

**Support Vector Machines**

This is one of the linear classifiers. This is very useful in web development document classification. If we are given different web documents which contain a very large number of features. Relatively low number of labelled training samples available. In such situations, employing boosted decision trees may not be a good idea. Usually works well until around 4k features. SVM is known to perform very well even with very low training data. Training scales well with a large number of features. SVM searches for the hyperplane with the largest margin, i.e., maximum marginal hyperplane using constrained convex quadratic optimization.

**NLTK Library & PySpark**

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language. We are planning to use this library along PySpark to tackle this big data problem to classify the tweets as either positive, negative or neutral.

*III. References*

1. Garbade, M. J. (2018, October 15). A Simple Introduction to Natural Language Processing. Retrieved April 26, 2020, from <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>
2. Algorithmia. (2020, March 30). What is Natural Language Processing? Introduction to NLP. Retrieved April 26, 2020, from https://algorithmia.com/blog/introduction-natural-language-processing-nlp
3. Chirag. (2017, June 21). NLP For Big Data Analytics - A Guide. Retrieved April 26, 2020, from <https://huddle.eurostarsoftwaretesting.com/nlp-for-big-data-how-nlp-will-revolutionise-big-data-analytics/>
4. Councill, I. G., McDonald, R. & Velikovich, L. (2010). What’s great and what’s not: learning to classify the scope of negation for improved sentiment analysis. In Proceedings of the workshop on negation and speculation in natural language processing (pp. 51–59). NeSp-NLP ’10. Uppsala, Sweden: Association for Computational Linguistics.
5. Kiritchenko, S., Zhu, X. & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, 50, 723–76