**49-733 Designing Smart Systems**

Final Project Spring 2018 April 25, 2018 Due May 2, 2018

**Twitter Geo-Sentiment Analysis Tool**

**Introduction**

**What is sentiment analysis?**

Sentiment is defined as "an attitude, thought, or judgment prompted by feeling."[1]. Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention. Sentiment analysis is one of the fastest growing research areas in computer science [2] and is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.

*Motivation : What problem would it solve? \*\**

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world [3]. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market. The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election [3].

**System Description**

This system being proposed uses natural language processing and machine learning classification algorithms to estimate and visualize  sentiment  for tweets. Tweets are short text snippets posted on Twitter, an online social network that allows users to upload short text messages “tweets” of up to 280 characters. The limited number of characters encourages users to construct focused and timely updates which makes tweets a perfect material to detect sentiment.

The goal of the system is to detect and visualize basic emotional properties embodied in the tweet, together with a measure of the confidence in the estimates in order to gain an overview of the wider public opinion behind certain topics. The geographical location of the tweets will also be identified and shown on a tweet map.

*Who benefits from the problem being solved?*

Companies interested to know what customers think of their products. Presidential campaigns wanting to gauge the popularity of their candidate and target certain groups of voters in specific states or geographical locations. Research institution interested in collecting opinion data and amatures who just want to play with this awesome tool!

*Competitive Analysis : How does the product or service solve the problem?*

**Out of the box libraries**

Many projects dealing with twitter sentiment analysis have relied mainly on out-of-the-box libraries such as “Text Blob” to determine the sentiment polarity and subjectivity of a tweet. While using such libraries makes it easier to detect sentiment, it can be less accurate. In this proposal, the goal is build sentiment dictionary using a combination of available sentiment dictionaries developed by researchers working on emotions and moods. Another drawback to out of the box libraries is that they steal the joy of getting your own data and training the model the way you please.

**Train your own model**

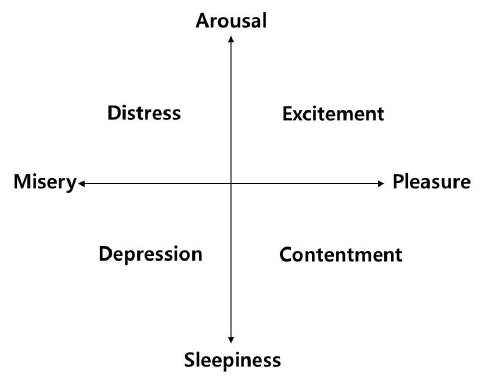
Some natural language processing libraries in python (such as TextBlob) allows programmers to train their classifier using data. By simply feeding a list of tuples to the classifier where the first item of the tuple is the text (or tweet) and the second being the assigned label (in this case sentiment).

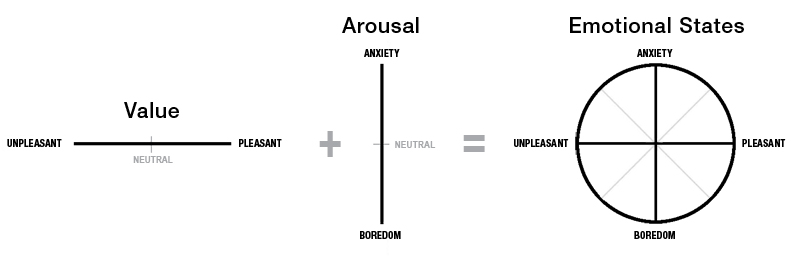
**Build your own classifier from scratch**

In many specialized domains where the sentiment needs to be determined with higher certainty or perhaps needs to be classified into more nuanced categories, a simple out-of-the-box sentiment classifier won’t be enough, and that’s where writing a specialized sentiment analysis classifier from scratch becomes necessary. Writing a classifier from scratch allows you more freedom to tinker with whatever machine learning classification algorithm that suits the problem.

**Sentiment categories**

Most sentiment analysis projects classified tweets into one of the three categories: Positive, negative or neutral. In this system, emotional models are proposed to produce more nuanced sentiment categories. These models often use emotional dimensions to position emotions on a 2D plane known as Russel’s emotional model [4] (see figures below).





In this model tweets will be displayed on an emotional scatter plot with pleasure and arousal as the horizontal and vertical axes respectively.

**Confidence Level of sentiment analysis**

In addition to detecting tweets’ sentiment, confidence level of the sentiment is also captured and visually represented (radius of circle or length of bar)

**Other interesting features**

1. Similarity analysis: Distances between tweets is dependent on text similarity (the closer the circles, the stronger the similarity)
2. Tag cloud generator of words associated with the topic of the tweet
3. Show trendy topics on twitter based on the user’s location
4. By clicking on a node representing the location of the tweet, the actual tweet pops up together with the twitter account and the sentiment
5. Time line vs. tweet count. (bar range = 15 min)

• *How does the solution decompose the problem into component subtasks and what subtasks are being solved by the computer and what by the computer?*

**Functions**

Get tweets from API

Analyze sentiment

Visualize output

**Files**

Twitter search: Capture Tweets from Twitter API given the topic and the geographic location

Preprocessing tweets : tokenizing, stemming etc.

TextBlob implementation: polarity and subjectivity of tweet

Implement more nuanced sentiment categories

Build specialized sentiment dictionary:

Classifying algorithms: Naïve Bayes

**Objects**

N/A

* **Algorithmic Plan** [5 pts]: A detailed algorithmic plan for how you will approach the trickiest part of the project.

SUBTASKS:

1. Data collection
   * The most challenging tasks would be:

* Cleaning data
* Building a good sentiment dictionary

1. Experiment with existing sentiment analysis library
2. Explore sentiment dictionaries
3. Algorithm selection
4. Testing and fine tuning

**Challenges**

1. Frequent use of slang and non-standard English
2. Use of sarcasm
3. Emojis and mixed emotions
4. Only few people choose to geotag their tweets

**References**

[1] https://www.merriam-webster.com/dictionary/sentiment

[2] The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers

[3] https://www.brandwatch.com/blog/understanding-sentiment-analysis/

[4] James Russell – A Circumplex model of affect