**Project Proposal**

* **Project Topic:** Twitter US Airline Sentiment Analysis.
* **Dataset:** ~15000 tweets containing following fields like, tweet text, sentiment of the tweets (positive, negative or neutral), airline name, tweet location, time, etc.

**We have chosen these two as a research ground for our papers:**

1. ***Sentiment Analysis of Twitter Data*** (cs.columbia.edu/~julia/papers/Agarwaletal11.pdf)

* **Key content and interesting ideas**: The major contributions of this paper are POS (part of speech)-specific prior polarity features and the use of a tree kernel to obviate the need for tedious feature engineering. The new features and the tree kernel perform approximately at the same level, both outperforming the state-of-the-art baseline.
* **Relate to course:** There has been a rapid increase in the use of social networking websites in the last few years. People most conveniently express their views and opinions on a wide array of topics via such websites. Sentiment analysis of such data which comprises of people's views is very important in order to gauge public opinion on a particular topic of interest. This paper proposes a method that can be used for sentiment analysis of tweets. In order to adapt these techniques for sentiment analysis of data procured from one of the social networking websites, Twitter, a number of issues and challenges need to be addressed, which are put forward in this paper. Also, this paper presents research topics in social media mining and addresses relevant theoretical foundations, methods, and tools**.**
* **Strengths, weakness, improvement:** The major strength of this paper is the data-set used. Unlike the other previously used data-sets, the tweets are collected in a streaming fashion and, therefore, represent a true sample of actual tweets in terms of language use and content. In this paper, two kinds of models are being investigated, tree kernel and feature based models and these models outperform the unigram baseline. There is room for more exploration on richer linguistic analysis, for example, parsing, and topic modeling.

1. ***An Approach to Sentiment Analysis: The Case Study of Airline Quality Rating***

(pacis-net.org/file/2014/1836.pdf)

* **Key content and interesting ideas**: To Do (Keval)
* **Relate to course:** To Do (Keval)
* **Strengths, weakness, improvement:** To Do (Keval)

**Proposed Work**

1. **What is your research question? Clearly define the research problem/question.**

* The research question in our project is: Analyzing the tweet sentiments of travelers who flew on various U.S. airlines in the month of February 2015. We want to answers various questions, like:
  + Text analysis of the tweets to find out the main reason behind the traveler’s positive/neutral/negative sentiments. Like if the tweet sentiment is negative, what were the travelers discussing, and when the tweet sentiment is positive, what were the travelers happy about?
  + A language model for analyzing the sentiments.
  + Find out which airlines which provide best and worst customer satisfaction.
  + Find out which states have most flyers and are most satisfied/dissatisfied.
  + Get the most discussed topics among various airlines and so on.

1. **Why is this an interesting question to ask and why would we care about the answer to this question or a solution to the problem?**

* This is an interesting question because; every large firm wants to keep their customers as happy as possible by providing them with the best service, so that the customer continues to be their loyal customer. The problem that we’re dealing here in this project is concerned with the analysis of users’ tweets sentiment about the quality of service of the airlines. By doing such analysis we would get some insight why the traveler was dissatisfied with a particular airline. The airlines can then use output of these analyses, rectify the problems and meets customers’ needs and eventually provide them with the better experience.

1. **Has any existing research work tried to answer the same or a similar question, and if so, what is still unknown?**

* There have been various works on the general domain of Twitter sentiment analysis, but there hasn’t been much work related specifically to analysis of sentiments on airline data. We found one paper related to this topic: pacis-net.org/file/2014/1836.pdf. In this paper, the authors are trying to find AQR (Airline Quality Rating) by comparing the tweets against a predefined corpus of subjective words. Also they have used CTM to model sentiment topic recognition. Their work in our opinion is very straightforward and can be improved. We can do some more thorough analysis of the tweets to get more intuitions and apply some more sophisticated algorithm to improve the model.

1. **How do you plan to work out the answer to the question?**

* We plan on doing following steps to answer the questions:
  + **Data Cleaning**: First of all we will clean the noisy data from the dataset manually and using tools like Google Refine.
  + **Preliminary Exploration**: Next we will do some preliminary exploration analysis of the data to find some interesting patterns and visuals that will give us some clues about the next steps. We will carry out this step using tools like Tableau, and libraries like PyPlot, matplotlib, etc.
  + **Sentiment Analysis**: Finally we will build some language processing model to analyze the tweets and find the reason behind the negative as well as positive sentiments, using NLTK package.

1. **How would you evaluate your solution? That is, how do you plan to demonstrate that your solution/answer is good or is reasonable?**

* To evaluate our solution, we would:
  + Compare our results with results of other similar paper/work.
  + Manually cross validate output with the ground truth.
  + Gain high accuracy.
  + Show the visualizations in a way that are in-depth and easy to understand.

1. **A rough timeline to show when you expect to finish what. List a couple of milestones.**

* Our expected timeline is as follows:
  + Milestone 1 (March 10th): Data cleaning and exploratory analysis.
  + Milestone 2 (April 1st): Sentiment analysis.
  + Milestone 3 (April 15th): Project completion.