**CSML1010 – Project Proposal – Group 11**

*Twitter Airline Sentiment*

**Background**

For our course project we have chosen to conduct a sentiment analysis on a data set containing approximately 14.5 thousand tweets pertaining to 6 major US airlines. Going into this project we knew that we were interested in choosing a data set and problem that closely aligned to solving a topical business problem. We also wanted to pick a data set that would offer a certain degree of challenge and learning opportunities.

Our chosen data set aligns well with these goals for several reasons. Firstly, in the dozen or so years that Twitter has existed it has contributed to a significant shift in the way that companies interact with their customers, primarily from a customer service point of view, but additionally in terms of PR, marketing and even logistics. One tweet from the right (or wrong) person can set off a landslide of responses that can quickly become out of control. This is a particular concern for the US Airline industry. Over the past several years there have been several high-profile incidents causing negative public sentiment towards US airlines. We believe that a robust sentiment analysis model – specifically focused on identifying these negative sentiments before they become a larger problem - could help airlines better manage their PR crisis response and customer service. A sentiment analysis model tuned to identify negative tweets with a high degree of accuracy could help airlines analyse and learn from past Twitter trends and to rapidly identify new ones as they are occurring.

Secondly, from the perspective of data science students a Twitter data offers an interesting challenge in terms of cleaning, interpretation and prediction. As Twitter caps each message at a short character limit, complete English is rarely used. The data set is full of abbreviations, slang, emojis and Twitter functions such as hashtags, mentions and re-tweets. Cleaning these out while retaining the information in the original tweet will be important to developing an effective model. As with many customer service data sets, this one is also likely to be unbalanced towards the negative sentiment side. Though this aligns well with our goal to primarily identify negative sentiment tweets, this will be important to consider as we clean the data.

**Data Cleaning**

**Data Exploration**