TWITTER AIRLINE SENTIMENT ANALYSIS

Introduction:

Airline service companies must interpret a substantial amount of customer feedback about their products and services. However, conventional methods to collect customers' feedback for airline service companies is to investigate through distributing and collecting questionnaires, which is time consuming and inaccurate. It needs labour to distribute and collect questionnaires to customers and also it will take too much effort to record and file those questionnaires considering how many passengers take flights every day. Beyond that, not all customers take questionnaires seriously and many customers just fill them in randomly and all of this brings noisy data into sentiment analysis. Unlike investigation questionnaires, twitter is a much better data source for sentiment classification for feedbacks of airline services. Because of the Big Data technologies, it has become very easy to collect millions of tweets and implement data analysis on them. This has saved a lot of labour costs which questionnaire investigations need. More than that, people post their genuine feelings on Twitter, which makes the information more accurate than investigation questionnaires. The other limitations for questionnaire investigations are that the questions on questionnaires are all set and it is hard to reveal the information which questionnaires do not cover.

As a result, sentiment analysis has become very popular in recent years for automatic customer satisfaction analysis of online services. Sentiment analysis is a sub domain of data mining, which are exploited to analyse large-scale data to reveal hidden information. Obviously, the advantages of automatic analysis of massive datasets make sentiment analysis preferable for airline companies.

Sentiment classification techniques can help researchers and decision makers in airline companies better understand customer feedback and satisfaction. Researchers and decision makers can utilize these techniques to automatically classify customers' 2 feedback on micro-blogging platforms like Twitter. Business analysis applications can be developed from these techniques as well.

There have been much research on text classification and sentiment classification, but there has been little on Twitter sentiment classification about airline services. Except applying popular sentiment classification approaches to tweets on airline services domain, it is also desirable to develop a new approach to further improve the classification accuracy

Objective:

Twitter is a really good source to get customers' feedback and marketing information in airline services, but there has been no perfect solution to automatically classify the massive amount of tweets, which leaves room for doing research in this area. A sentiment analysis job about the problems of each major U.S. airline. Planning on scraping Twitter data for a random month of 2015 and to classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

I intend on answering some questions, like:

- Text analysis of the user tweets to find out the reasons behind the user's sentiments.
- Find out which airlines which provide best and worst customer satisfaction.
- Get the most discussed topics among various airlines.
- Reasons for negative sentiment (and by airlines).
- Top negative and positive words.
- Other interesting stats and graphs.

For implementation purposes I would like to use R and Python (Text classification) to infer about the dataset.

Timeline:

Below is the tentative chronology of events as per my knowledge and experience. Hoping for everything going as per the timetable. Worst case estimate can be + 1 week.

Milestone	Duration
Data Collection	1 week
Data Pre-processing	2-3 weeks
Querying	2 weeks
EDA	2 weeks
Visualization	2 weeks
Final Project	1 week

Limitations:

There are might be certain limitations to this project as per my knowledge. First of all, the tweets collected from the Twitter API are not as pure as required for the sentiment classification. By searching tweets with keywords "flight" and the airline brand, there is a probability that 30% to 40% of dataset might be irrelevant tweets. Airline companies still need to further improve the accuracy of tweet data retrieval. Secondly, compared to the tweet data existing in Twitter, the dataset collected and used in this project is a very tiny part. This is a problem to solve for doing scale sentiment classification by applying big data techniques. However, it requires expensive infrastructure investment to do this kind research and application. Besides that, my tweet data are very messy and they contain a lot of typos and abbreviations. Even though adopting feature selection techniques to reduce the dimensionality, it still cannot group all words with the same roots into one stem. It is desirable to auto correct all the typos and to extend the abbreviations to regular words, which requires high level Natural Language Processing techniques. More than that, the balanced class distribution is not a real world case and it might cause over-fitting problem in positive class. In the future, more complicated models are expected to be built to solve this problem.

Moreover, for different airline brands, the features for sentiment classification might be different from each other and it is valuable to train sentiment classification models for different airline brands. To give more specific and valuable information for the decision makers of the airline companies, sentiment classification can be applied to the tweets about their airline services and produce detailed reports. Last but not least, there are also many further research directions, which can be worked on. Other information like the users who tweet them, the times of the retweets and other factors are also potentially useful. The time series analysis of the twitter sentiment about airline services is also an interesting topic and the time data is available in the tweets retrieved from the Twitter API.

Ethics of Social Media Research:

There are a number of ethical considerations to keep in mind when using social media data for this research. One of the key concerns hinges on the extent to which social media data should be treated as public versus private data. Even though social media data are publicly available, social media users may not intend or wish for their data to be used for research. Users may not be aware that their social media data is publicly available and may have expectations of privacy even in public settings. The distinction between public and private data becomes additionally complicated by the fact that machine learning algorithms can make inferences about private attributes, even if not explicitly stated in public data. These issues have been properly addressed in this research and no user information has been neither been collected nor has been used for analysis and inference.