An approach: Airlines service feedback using Sentiment Analysis

CS 6675 – Advance Data Science

What is the key technical content and interesting ideas behind the papers?

The paper has focused on developing a sentiment classification system for airlines services using the feedback provided by the passengers using Twitter. It is very difficult to gather passengers feedback taken by airlines using online surveys and questionnaires. The Social media has provided the customers a platform where everyone can put their thoughts and views. The authors of the paper have considered Twitter as the most important means of collecting feedbacks of airlines. In the paper, a collaborative sentiment analysis strategy was applied based on Majority Vote principle of 6 individual classification approaches. They trained and tested the system using nearly 13k tweets. The method developed by them outperformed the accuracy of all the 6 algorithms individually.

How do the papers relate to the topics presented in the course?

Today the social network is a very important part of our life. People share their views about anything starting from government to different products of different brands using the social media so that other users can get aware. Sentiment Analysis system can be created to easily filter out positive, negative, and neutral feedbacks provided by the users. The author of the paper has developed a collaborative approach and compared it with 6 popular classification methods. The paper uses the technique of social media mining and other theoretical concepts to solve the problem.

What are the strengths of the paper? What are its weaknesses? And how might you go about improving on the weaknesses?

Strengths:

- Recent researches have been evaluated and considered before developing a new approach
- They considered Twitter Sentiment analysis along with conventional sentiment analysis.
- Compared Information gained for features.
- Used various evaluation techniques for comparing sentiment analysis.

Weakness:

- Considered less than 13k tweets.
- Considered only text of tweets, but the retweets can also be a useful feature.
- Haven't used NLTK trainer
- Initially they manually labelled the tweets with the polarity.

Improvements:

- We are planning to collect more than 20,000 tweets.
- We will be using more than 2 keywords in Twitter Search API, to collect more relevant service related tweets.
- We would be using the Textblob for calculating polarity of each tweets.

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

The research addresses analyzing the feedback of customer's who flew on different U.S. airlines. We were planning to give the answers to the following questions:

- Content investigation on the feedback of the tweets to discover the principle explanation for the customer's positive/impartial/negative conclusions. Like if the tweet opinion is negative, what were the customer talking about, and when the tweet estimation is certain, what were the customer upbeat about?
- Which airlines provides the best customer satisfaction and which provides the worst?
- Discover which states have most flyers and are most satisfied/disappointed.

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?

There are 2 reasons for finding this topic as interesting:

Today, lot of people travel everywhere (domestic or abroad) using airlines. Usually it is seen that passengers are not happy with the services provided by the airlines. They provide feedback or their complaints using the social media like Facebook, Twitter, and even to the feedback survey of the airlines. But the feedback survey results are not available to other passengers. In such scenarios, social media can play a very important role. On collecting feedbacks from social media sites a very good and efficient review or rating system can be formed for the benefit of the customers.

Researchers and decision makers can utilize these techniques to automatically collect customer's opinions about airline services from various micro-blogging platforms like Twitter. Business

analysis applications can be developed from these techniques as well. There has been much research on text classification and sentiment classification, but there is little research done directly linking to Twitter sentiment analysis about airline services

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

A lot of work has been done on text mining which includes information retrieval, lexical analysis etc., but use of special characters like @ and # make it very difficult. Earlier works didn't consider Neutral Sentiments. Very little work has been done on twitter sentiment classifications and even less work has been done for airline services. Using various classification techniques, we will do thorough analysis of the airline tweets to get more insights and apply some algorithm to improve the model

There have been different papers who tries to take a shot at the general area of Airplane Quality, however there hasn't been any research specifically dedicated for investigation on customer's feedback on airline services.

How do you plan to work out the answer to the question? (At the proposal stage, you are only expected to have a sketch of your methods

1. We are planning to get the data from kaggle and Twitter.
2. Then we are planning to perform data cleaning and data pre-processing techniques.
3. Then we will be working on processing the cleaned tweets using tokenization and removing of special characters and links.
4. Using these tweet text for polarity calculation and labelling the data.
5. We will be doing various analysis and visualizations on the sentiment and tweets.

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

For evaluating our solution, we will:

- Compare and draw inferences of our results with results of referred paper.
- Visually cross validate paper output with our output.
- We are planning to use some current techniques to increase the accuracy.

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

20th March: Data collection, Data cleaning and exploratory analysis.
5th April: Build our model and perform Sentiment analysis
20th April: Visualization, Report generation using Latex.

References:

1. https://www.kaggle.com/crowdflower/twitter-airline-sentiment
2. Main Research paper - http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7395820&tag=1
3. Secondary Research Paper - https://www.researchgate.net/profile/Keng_Siau/publication/265381265_An_Approach_to_Sentiment_Analysis_-_The_Case_of_Airline_Quality_Rating/links/554d27880cf29f836c9cd7f0.pdf