CURRENT TREND OF AIRLINES USING TWITTER DATA

Harshitha Yerraguntla
Computer Engineering, San Jose State University, California

Abstract:

With the increase in the public and privately available information, many opinions are available on social media, blogs, forums in the form of reviews, tweets, feedback. Data is available in the form of unstructured or semi-structured data. By using the sentiment analysis, we are transforming unstructured data of the public opinions about airline's customer service to structured data. This can be further used for commercial applications like market analysis, product feedback, product service, product reviews, public relations, etc. In this project, we are concentrating on polarity classification which classifies a sentence as a positive, negative or neutral opinion. People generally choose the airlines that have many people reviewed positively. Sentiment analysis helps you to analyze your feelings or opinions in the tweet. Businesses can use sentiment analysis to improve business quality. Sentiment analysis is the subfield of natural language processing which is the part of Artificial intelligence, where the machines process the human languages. Identifying if the user opinion is negative, positive, or neutral is polarity classification. Airlines executives can use the supervised learning algorithm to observe the sentiment associated with their airlines and the impact of sentiment due to various campaigns introduced. In addition to polarity, we have identified the reasons for polarity using topic modelling as well. This will tell the Airlines, the reason for their sentiment and they can take necessary action to improve the sentiment.

1. INTRODUCTION

According to Wikipedia, Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials [1]. Most of the people use social media for representation of their views, opinions, ideas, and sentiments. They share their opinions related to services on social media. The best way to understand the people opinions on a product/service is by collecting the tweets related to that product. It's humanly impossible to read all the tweets, so we need a system that can summarize the contents of these tweets. This data on social media can be scraped to gather the views of the customers on services provided. Based on available data, we are developing an accurate multiclass classifier which scrapes the current data from the twitter and gives the sentiment or polarity associated with the airlines. In addition to predicting the sentiment, we have given the reason for the sentiment. This solution helps the executives of the airline company to continuously monitor the sentiment of the airlines and to take necessary actions for improving the services and facilities which in turn increases business.

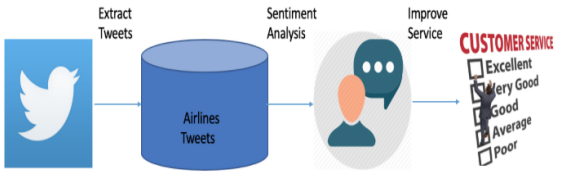


Figure 1.1 Workflow Diagram

2. WEB SCRAPING

Twitter is an amazing microblogging website where users express their opinions using messages called as tweets. If a person has tweeted a message, the same message can be retweeted by many other users as well. For example, if a person tweets something about an Airline, the tweet will be retweeted by many other users and the message will spread quite quickly in social media. The spreading to the message/sentiment is even high if any famous personality is involved in the tweet. This data/tweets can be used for social media analysis. Since we are concentrating on the current trends which is relevant to the Airlines business, we require most recent data to estimate the current trends associated with the airlines. After requesting the developer account from twitter, current

data is scraped from twitter to know the current trends for different airlines. Using twitter API, past one week's tweets can be downloaded. Tweets are downloaded based on the handles. These are the steps to download the data from Twitter.

1. Request Twitter for a developer account to download any data form twitter API
2. Created an app in twitter and generate the for keys to download data. The four keys are:
   - Consumer API Key
   - Consumer API secret key
   - Access token
   - Access token secret key
3. Using the above generated API access and token keys and tweepy library in Python, we are able to access the data from twitter
4. Based on the twitter handle associated with airlines, the appropriate tweets are downloaded in JSON format
5. The downloaded tweets in JSON are converted to CSV for sentiment Analysis

3. DATASET INFORMATION

In the training data there are nearly 15000 tweets belonging to six airlines are labeled manually as positive, negative or neutral. Totally 17% of the data labeled as positive, 62% data labeled as negative, 21% of data labeled as neutral based on reviews. Figure 3.1 shows the graphical representation of the distribution of sentiments. This labeled data is used to train the model which will help us to know the latest sentiment of airlines. The data consists of 15 attributes such as tweet text, airline name, airline sentiment, tweet id, negative reason, negative reason confidence.

Since the problem goal is to classify the current tweets into positive, negative, neutral groups with respect to airlines, only required attributes such as Tweet text, airline sentiment, airline name is considered into the account of attributes. Current tweets are scraped from the twitter using developer APIs.

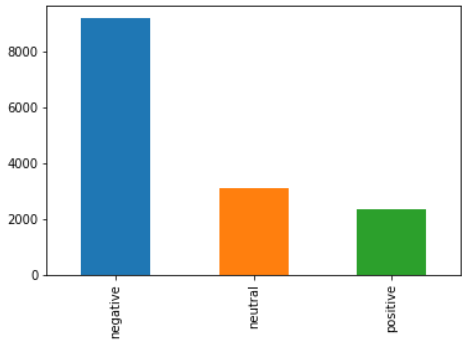


Figure 3.1 distribution of sentiments

There are 6 Airlines information in the training dataset. They are:

- Southwest Airlines
- Virgin America
- US Airways
- United Airlines
- Delta airlines
- American Airlines

Sentiment associated with each of the Airlines is shown in the figure 3.2

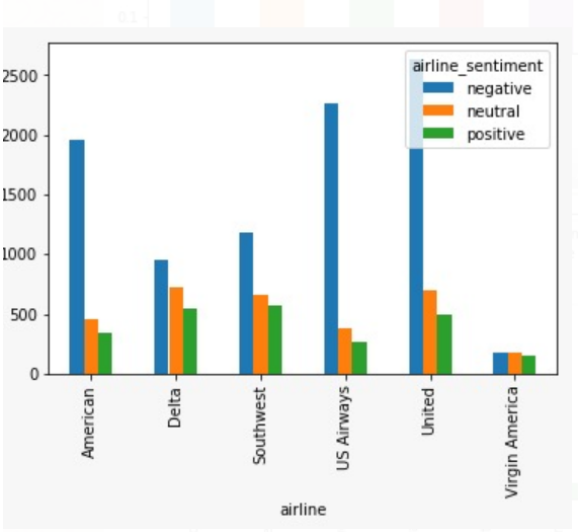


Figure 3.2 sentiment by each Airline

The following figure 3.3 showed the number of tweets associated with each of the 6 Airlines. United has highest number of tweets and Virgin America has the least number.

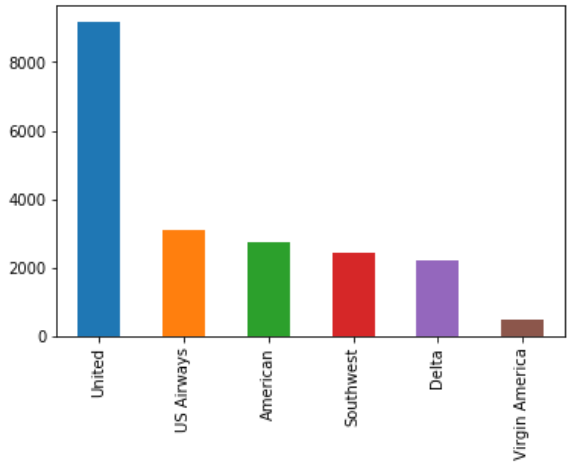


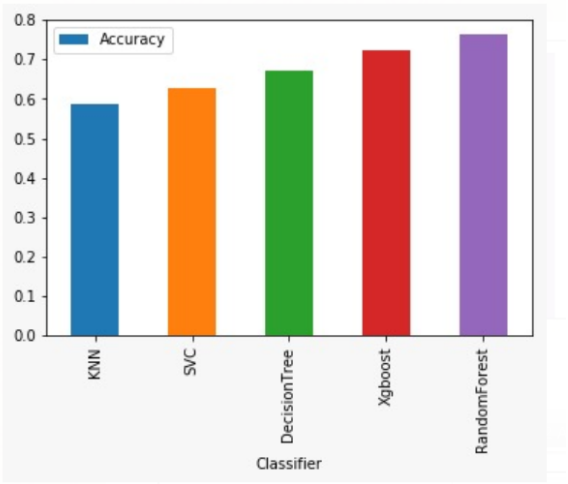
Figure 3.3 number of tweets for each Airline

4. IMPLEMENTATION

As part of implementation, the first thing to do is data preprocessing. As part of preprocessing the following things were performed.

- In each of the tweets English stop words are removed.
- Words less than 3 characters are removed
- Punctuations and special characters are removed
- Numbers and characters not part of English language is removed
- Tweets are tokenized before sending them to NLTK

Using the various visualization techniques, we can infer the most common issues of the customer with respect to airlines and identify unusual trends in case of any issues such as unlabeled data. Most of the tweets are from the people who are unsatisfied with the airline's service. The most negative words in the tweets help the management to improve the airline's facilities/services with respect to the majority tweets. We perform the sentiment analysis on the tweets giving a score to each of them. Used NLTK package in Python that helps to work on human languages. Various machine learning models such as Decision Trees, Random Forests, KNN, SVM, xgboost are used to predict the sentiment associated with each tweet. K-fold Cross-validation technique is used for tuning the parameters. Regularization techniques are used to avoid overfitting of the model. Random Forest gave the best accuracy of all the models and with xgboost coming next best. Please see Figure 4.1 to see the accuracy of the various models being tried. We have chosen Random Forest to predict the sentiment on the new/unseen data.



4.1 Different model accuracies

Also, after predicting the sentiment for the most recent tweets, some of the tweets were manually examined to see how good our algorithm is predicting on unseen data. After manually examining the tweets, it looks like our model is predicting quite accurately. Please find the sample positive, neutral and negative comment predicted from the recent tweets below:

Positive Tweet: b'Thank you Michael and Diana for helping our family @SouthwestAir you redeemed yourselves #thankyou'

Neutral Tweet: b'@SouthwestAir Good bye Austin hello Dallas!! #home #needanap'

Negative Tweet: b'@SouthwestAir Lost luggage and still no update stuck in Florida no clothes granddaughter cannot perform in competition \xf0\x9f\x98\xad'

5. RESULTS AND OBSERVATIONS

There are a greater number of negative tweets compared to positive tweets. By this we can observe that the customers who are not satisfied by the service,

are the ones who tweets more frequently on the twitter. Table 5.1 shows the sentiment associated with total 10 airlines in last one week. There are very a smaller number of tweets for US Airways, SkyWest, Spirit and Virgin America. It is difficult to conclude the sentiment for these four airlines given that the number of tweets are too less.

Coming to the neutral tweets, there are more neutral tweets for the southwest Airlines and JetBlue. By this analysis, as of now, southwest is the best airlines compared to the other five airlines. When it comes to positive sentiment as well Southwest Airlines and JetBlue has highest proportion of the total tweets. When it comes to negative sentiment American Airlines, Hawain Airlines, United Airlines and Delta Airlines have the highest negative sentiment. These Airlines must see the reason for such a high negative sentiment and address the customers issues to have a positive sentiment.

	Positive	Neutral	Negative	% of -ve comments	% of +ve comments
Alaska Air	86	2	912	91.20	8.60
American Airlines	180	14	3106	**94.12**	5.45
Delta	198	28	3074	**93.15**	6.00
Hawaian	38	8	730	**94.07**	4.90
JetBlue	114	164	722	72.20	**11.40**
Skywest	3	1	18	81.82	13.64
Southwest	361	84	2855	86.52	**10.94**
Spirit	71	13	916	91.60	7.10
US Airways	3	1	31	88.57	8.57
Virigin America	14	29	81	65.32	11.29
United	178	23	3099	**93.91**	5.39

Figure 5.1 Sentiment of different Airlines during last one week

We investigated the reason for negative sentiment for United Airlines and the following are the main reasons:

- Cancelled Flight
- Late Flight
- Bad Flight
- Flight Booking Problems
- Customer Service Issue
- Lost Luggage
- Long Lines
- Flight attendant complaints
- Damaged Luggage

Please see the Figure 5.3 to see the summary of reasons for Negative sentiment for United Airlines.

There are bunch of reasons for negative sentiment which are not related to Airlines at all. The reasons like Damaged Luggage and Lost Luggage are not all related to the Airlines. These reasons are more to do with the Airport. Airlines can use the tweet location to identify at which of the airports their customers are facing issues due to luggage and take corrective measures by working with Airport authorities.

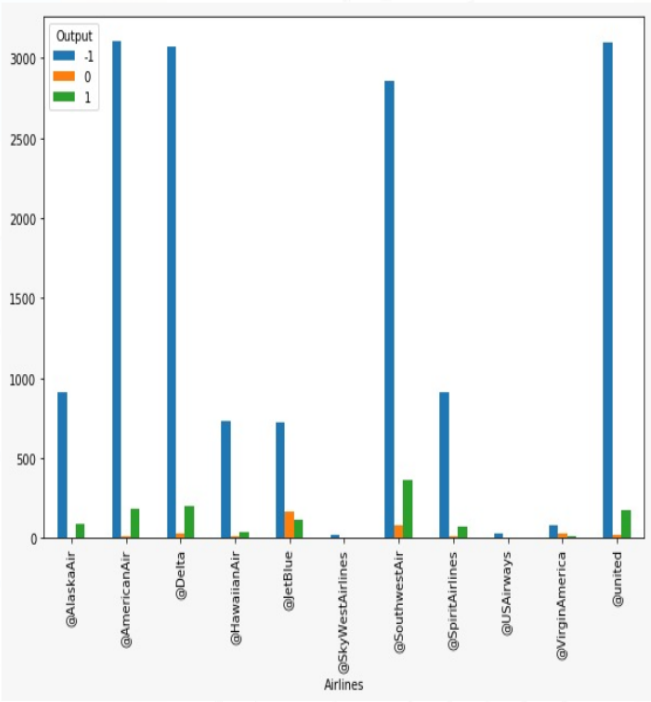


Figure 5.2 Sentiment associated with each of the ten airlines in last one week

There are few other reasons for the negative sentiment for United Airlines, which are related to Airlines service. These are issues like Customer Service issue, Long Lines, cancelled flights and flight attendant complaints. These can be taken care by Airlines management as these things are mostly within their control. They can increase the number of counters where there are complaints regarding long lines. Similarly, they can train their Flight Attendants and

customer service representatives to improve customer satisfaction and that can improve their sentiment.

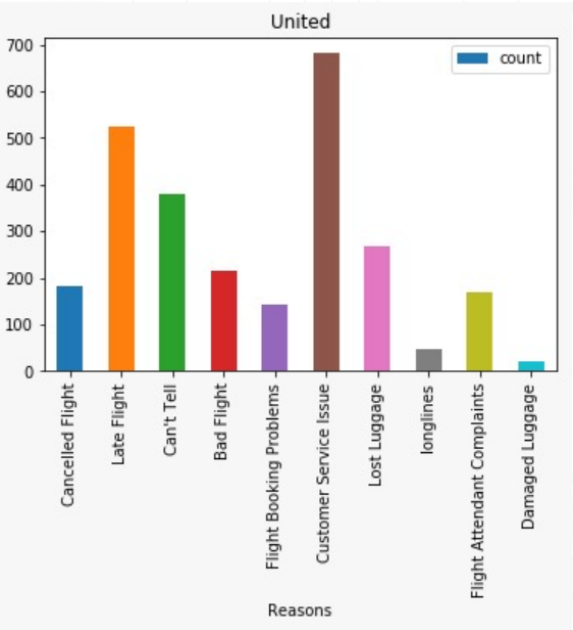


Figure 5.3 Reasons for Negative sentiment for United Airlines

Another important factor contributing to the negative sentiment of United Airlines is that their customers are facing Flight Booking problems. Given that most of their customers would be booking using their website or mobile App. It is time to revamp them to make more customer friendly to have a smooth way to book their tickets.

Conversely the positive sentiment for Airlines is associated with Customer Service in case of JetBlue and Southwest Airways.

6. CONCLUSION

Sentiment Analysis helps to reveal the opinions of the passengers on the airline services. More people prefer to communicate with the social media instead of word of mouth. Apart from sentiment, text mining will give us the information regarding the reason for the associated sentiment. In this case study the Negative sentiment is mainly due to poor Airline service, Airline's customer service and Issues with the luggage. Conversely positive sentiment is due to Customer support of the Airlines. For the Airlines executives to know the sentiment and reason for the sentiment, they don't need to go through all the tweets. They can just use text mining to get the essence of the customer feedback.

REFERENCES :

1. **https://en.wikipedia.org/wiki/Sentiment_analysis**
2. **http://cs229.stanford.edu/proj2016spr/report/042..Pdf.**
3. **https://www.kaggle.com/crowdflower/twitter-airline-sentiment**
4. **http://www.slideshare.net/ajayohri/twitter-analysis-by-kaify-rais**
5. **https://www.researchgate.net/publication/315643035_Online_Social_Media-based_Sentiment_Analysis_for_US_Airline_companies**
6. **https://www.semanticscholar.org/paper/An-Approach-to-Sentiment-Analysis-the-Case-of-Adeborna-Siau/dc8e272b9f56b935b926603bdf42eb033c6e94a3**