Twitter USA Airline Sentiment Analysis and Classification System

Sunil Pankaj EE14BTECH11032@iith.ac.in
Pratik Barule EE14BTECH11006@iith.ac.in

Abstract— **In airline service industry, it is difficult to collect data about customers' feedback by questionnaires, but Twitter provides a sound data source for them to do customer sentiment analysis. However, little research has been done in the domain of Twitter sentiment classification about airline services. In this paper, we analyse data and make prediction of classes using Naive Bayes Classifier.**

I. INTRODUCTION

Customer feedback analysis is very important for improving airline services. However, the conventional methods is to collect customers feedbacks through distributing, collecting and analyzing questionnaires, which is time consuming and often inaccurate because 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 various 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 the data. 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.

Sentiment classification techniques can help researchers and decision makers in airline companies to better understand customers feeling, opinions and satisfaction. Researchers and decision makers can utilize these techniques to automatically collect customers opinions about airline services from various 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 is little research done directly linking to Twitter sentiment analysis about airline services. Another issue is how to compare and improve Twitter sentiment classification accuracy among different classification methods.

II. RALATED WORK

Sentiment classification is a division of text mining, which includes information retrieval, lexical analysis and many other techniques The simplest way to do sentiment classification is using the Lexicon-based approach [1], which calculates the sum of the number of the positive sentiment words and the negative sentiment words appearing in the text file to determine the sentiment of the text file. The Naive Bayes method has been a very popular method in text categorization because its simplicity and efficiency [2].The theory behind is that the joint probability of two events can be used to predict the probability of one event given the occurrence of the other event. They key assumption of the Naive Bayesian method is that the attributes in classification are independent to each other, which considerably reduces the computing complexity of the classification algorithm.

Big Data social data analysis has been very popular [3]. Because Twitter provide public access to its streaming and historical data, it has become a very popular data source for sentiment analysis and many work has been done in this area. J.Read used emoticons, such as ":-)" and ":-(", to collect tweets with sentiments and to categorize them into positive tweets and negative tweet. They adopted Nave Bayesian approach and the Support Vector Machine approach, both of which reached accuracy up to 70% [4]. In the research of Wilson et al, they used hashtags to collect tweets as the training dataset. They tried to solve the problem of wide topic range of tweet data and proposed a universal method to produce training dataset for any topic in tweets [5]. In their experiments, it showed that training data with hashtags could train better classifiers than regular training data do. But in their research, the dataset were from libraries and they neglected the fact that tweets with hashtags are only a small part of real world tweets data. Pak and Paroubek proposed an approach, which can retrieve sentiment oriented tweets from the twitter API and classify their sentiment orientations [6]. From the test result, they found that the classifier using bigram features produces highest classification accuracy because it achieves a good balance between coverage and precision. But the data source is biased as well because they retrieved only the tweets with emoticons and neglected all other tweets that didnt contain emoticons, which are the majority of tweets. In this work, they didnt consider the existence of the neutral sentiment and classifying these tweets is very important for tweet sentiment analysis.

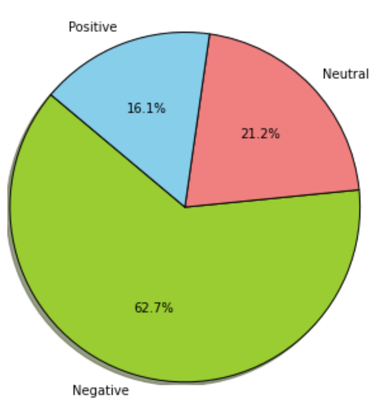
Little work has been done on twitter sentiment classifications about airline services. Conventional sentiment classification approaches, such as Nave Bayesian approach, have been applied to some tweet data and the performance was not bad [6]. Lee et al used twitter as the data source to analyze consumers communications about airline services [7]. They studied tweets from three airline brands: Malaysia Airlines, JetBlue Airlines and SouthWest Airlines. They adopted conventional text analysis methods in studying Twitter users interactions and provided advices to airline companies for micro-blogging campaign. In their research, they didnt adopt sentiment classification on tweets, which will be more salient for airline services companies to understand what customers are thinking. In the handbook of Mining Twitter for Airline Consumer Sentiment, Jeffery Oliver illustrates classifying tweets sentiment by applying sentimental lexicons [8]. This handbook suggests retrieving real time tweets from Twitter API with queries containing airline companies names. The sentiment lexicons in this method are not domain specific and there is no data training process or testing process. By matching each tweet with the positive word list and the negative word list, and assigning scores based on matching result to each tweet, they can be classified as positive or negative according to the summed scores. The accuracy is unknown since it is not considered in this book.In our work, more than 100,000 tweets are collected, and unigrams, bigrams, trigrams and the Information Gain algorithm are applied into feature selection, which is much less biased. Besides that, their work did not present details about the classification approaches and comprehensive evaluations.

III. About Data

We used twitter data provided by Kaggle.for this project twitter data was scraped from February of 2015. there is a dataset containing 14640 tweets in our work. In the dataset, 2363 tweets are labeled positive, 9178 tweets are labeled negative, 3099 tweets are labeled neutral. dataset contains 15 features: tweet-id, airline-sentiment, airline-sentiment-confidence, negative-reason, negative-reason-confidence, airline, airline-sentiment-gold, name, negative-reason-gold, retweet-count, text, tweet-coord, tweet-created, tweet-location,user-timezone.

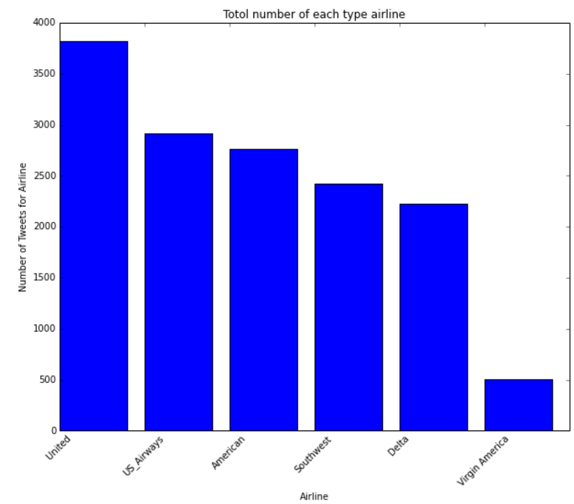
TABLE I
Class Distribution

Class	Positive	Negative	Neutral
Tweets	2363	9178	3099

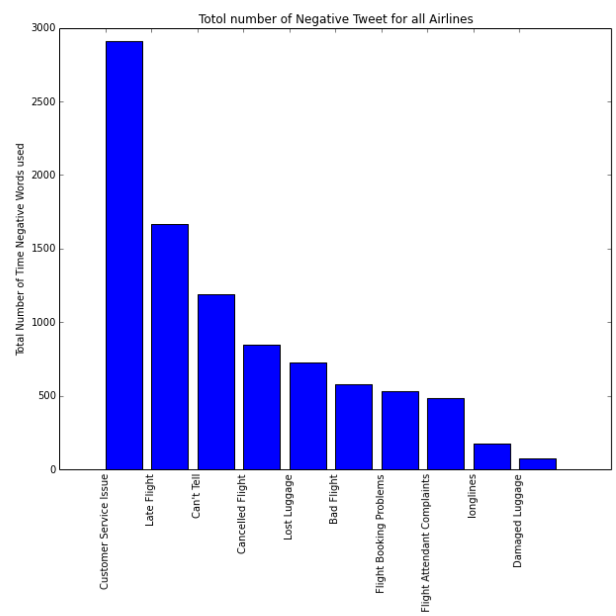


IV. Sentimental analysis

We tried to analyze sentiments for all the top US airlines differently. The tweets for each airline were considered separately. After dividing the tweets on the basis of the airline company they refer to we got the following distribution:



In a feedback negative comments are more important than positive comments. In machine learning terms negative comments provide more relevant data in order to improve the services.When we analyzed the our data for negative comments in order to decide which factors were causing more trouble to the passengers here is what we got:



The above graph can be used by airline companies to see the problems faced most by the people and it makes the analyzing part very simple.

The above graph is drawn from the word cloud show below. A word cloud is just a fancy way used to illustrate the words used. Following is the Word Cloud we got from the data for negative sentiment.

V. METHODOLOGY AND CLASSIFIER DESIGN

For model training and classification, balanced class distribution is very important to ensure the prior probabilities are not biased caused by the imbalanced class distribution.Unlike formal publications, the texts on social networks and blogs are unedited texts, which means they are not bound to strict grammar rules and the requirements of correct spelling. Typos and abbreviations happen a lot in social network postings, especially in tweets. To solve this problem, First we are removing @-usernam, emoticons, URLs, hash-tags, numbers, words with and without dashes and apostrophes. After that we are breaking the text down into words. Tokenisation is one of the most basic, yet most important, steps in text analysis. The purpose of tokenisation is to split a stream of text into smaller units called tokens, usually words or phrases. After getting token we adopt stemming techniques to stem the different reflections of the words to their word stem. For example, all of the different forms and reflections of the word ”cancel” such as ”cancelling”, ”cancelled” and ”canceled” can be converted to an identical stem word ”cancel” through stemming techniques and removing duplicated words. Nevertheless, stemming techniques still can considerably reduce the sparsity of the features. The next step is to remove the Stop-Words. Stop words are the word like the, that, an, there, am, herself etc. Now our data is clean. Next step is detokenize cleaned dataframe for vectorizing. We have used CountVectorizer to convert the data into a Sparce Matrix format. This sparce matrix then can be used as input to the machine learning algorithms. In this paper we have only consider only Naive Bayes classifier.

In sentiment classification, features can be unigrams, bigrams, trigrams and more. The reason for taking N-gram features from text documents is because N-gram features indicate different sentiment information than the unigrams do. Sometimes it is because the preceding word in an N-gram phrase is a negation, which can reverse the sentiment orientation of the unigrams in the phrase to the opposite sentiment orientation and give the N-gram phrase the opposite sentiment orientation to the unigrams in it. Every two consecutive words in a tweet document are considered a bigram. So for a tweet document with N unigrams, there are (N-1) bigrams for this tweet document. Every three consecutive words in a tweet document are considered a trigram. So for a tweet with N unigrams, there are (N-2) trigrams. Actually, we can consider even longer multi-grams in sentiment classification, such as four-grams or five-grams. However, there are several reasons for not doing that. First of all, it will make the transformed matrix even sparser and make the sentiment classification not implementable. Besides, as the length of the N-gram becomes longer, the N-gram features for each tweet document will be more distinct from the N-gram features from other tweet documents. There has been research that from bigrams to multi-grams, the Information Gain for each level of N-gram decreases as the length of the multi- grams increases [9]. As shown in Figure 2 the Information Gain decreases as the feature length increases.

A. Naive Bayesian Classifier

The Naive Bayesian method is one of the most widely used methods to classify text data. The Naive Bayesian algorithm assumes that the elements in dataset are independent from each other and their occurrences in different dataset indicate their relevance to certain data attributes. The Naive Bayesian classifier treats each tweet document as a bag-of-words. The Nave Bayesian classifier passes a single tweet document and calculates the products of the probabilities of every feature occurring in this tweet for each of the three sentiment orientations, positive, negative and neutral. The sentiment orientation of this tweet is classified to one of the three sentiment orientations, which gets the biggest probability product. In our work, we utilize the NaiveBayes algorithm provided in SKlearn to implement experiments and tests.

VI. EXPERIMENT AND EVALUATION

Accuracy using naive bayes classifier: 0.769672131148

(1) Accuracy Evaluation Based on F-measure

In accuracy evaluation of classification, there are Recall, Precision and F-measure to evaluate the overall accuracy of the classifier.

a) Recall

Recall is the fraction of the correctly classified instances for one class of the overall instances in this class [2].

TABLE II
RECALL

Class	negative	netural	postive
Recall	95%	38%	57%

b) Precision

Precision is the fraction of the correctly classified instances for one class of the overall instances which are classified to this class [2]

(c) F-measure

To get a comprehensive evaluation of the classification, F- measure is developed to integrate the Recall and the

TABLE III

PRECISION

Class	negative	netural	postive
Precision	78%	70%	80%

Precision. The F-measure can be expressed as

$$F_B = \frac{(1+B^2)*(precision*recall)}{(B^2+precision+recall)}$$

This is a general form of F-measure and the parameter is used to change the weights for Precision and Recall in calculating the F1-measure value. In our work, because recall and precision are equally important [2]. We set to 1, and it is called the harmonic mean of precision and recall.

TABLE IV

F1-MEASURE

Class	negative	netural	postive
F1-measure	86%	49%	66%

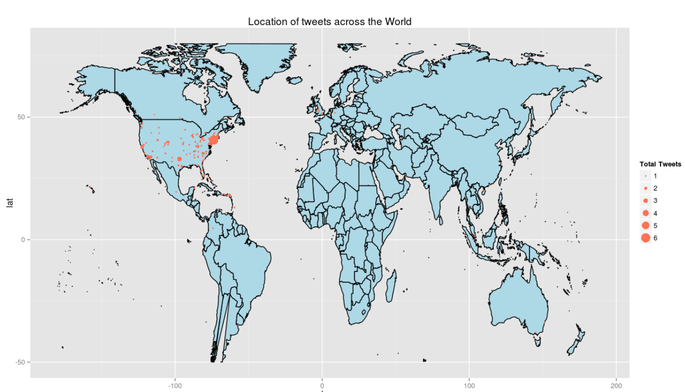
(d) Classification Report

Accuracy	precision	recall	f1-score	support
Negative	78%	95%	86%	2291
Neutral	70%	38%	49%	774
Positive	80%	57%	66%	595

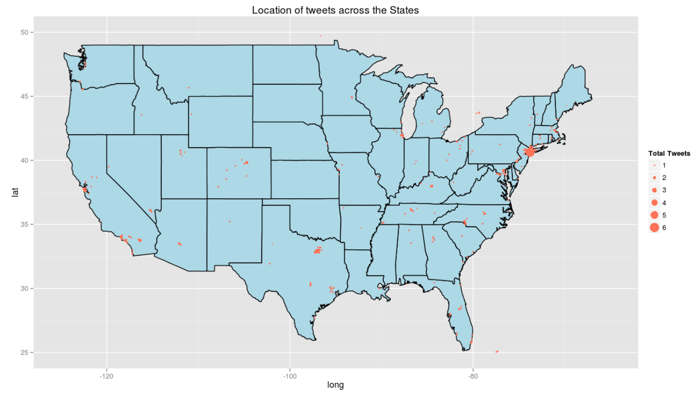
VII. CONCLUSION

In this paper we have tried to make our contribution to the area of analysis of twitter sentiments. In the domain of twitter sentiment analysis about airline services, little work has been done. For the airline services domain, the sentiment classification accuracy is high enough to implement customer satisfaction investigation. This approach is applicable for the airline companies to analyze the twitter data about their services. There is also much further research, which can be worked on. In our project, only the texts of the tweets are considered and other information like the users who tweet them, the times of the retweets and other factors are also potentially useful.

Following are the maps which tells us about the areas from which we are getting the most tweets.



As most of the tweets are coming from USA it's worth to analyze it separately



As expected most of the tweets from big cities like San Francisco, Las Vegas, New York etc.

Here is github link which contain everything (code, figures, data)related to project:

https://github.com/sunilpankaj/Twitter-US-Airline-Sentiment

VIII. REFERENCES

[1] Cambria, Erik, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Statistical approaches to concept-level sentiment analysis." IEEE Intelligent Systems 3 (2013): 6-9.

[2] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[3] E. Cambria, H. Wang, and B. White, Guest editorial: Big social data analysis, Knowledge-Based Systems, vol. 69, pp. 12, 2014.

[4] Read, Jonathon. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification." Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005.

[5] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005.

[6]] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[7] Dharmavaram Sreenivasan, Nirupama, Chei Sian Lee, and Dion Hoe- Lian Goh. "Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines." Program 46.1 (2012): 21- 42.

[8] Breen, Jeffrey Oliver. "Mining twitter for airline consumer sentiment." Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications (2012): 133.

[9] Cheng, Hong, et al. ”Discriminative frequent pattern analysis for effective classification.” Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.