A Comparison of Approaches to Twitter Sentiment Classification in Airline Services Domain

Anonymous First, Anonymous Second

Abstract. Twitter users often tweet their feelings about airline service experiences, which provides massive valuable data and information to airline services companies. Collecting and analyzing the sentiments from these tweets could be important for airline services companies to track customer satisfaction information and to discover salient marketing opportunities. We applied four sentiment classification approaches and different feature selection methods to real world tweets in airline service domain and compare their approaches include lexicon-based, SVM based, Naïve Bayesian based and Bayesian Network based. The results of our experiments demonstrate that the SVM approach works best in real-word practice on sentiment classification of tweet data.

Keywords: Twitter data mining, sentiment analysis, Airline services, classification comparison

1 Introduction

For airline services companies, it's crucial to comprehend customers' feedback about their products and services in substantial scale. However, conventional customer satisfaction analysis methods, like questionnaire investigation, are exceedingly time-consuming and the results are highly inaccurate because of sample errors. As a result, text sentiment analysis has been getting very popular in recent years for automatic customer satisfaction analysis of online services. Sentiment analysis is the application of data mining methods, which are exploited to analyze large-scale data to reveal hidden information. Obviously, the advantages of automatic analysis of massive datasets make sentiment analysis preferable for airline services companies.

In this study, we developed several sentiment classification algorithms including a lexicon-based classifier, a SVM classifier, a Naïve Bayesian classifier and a Bayesian Network classifier, and tested them with datasets retrieved from Twitter API. Based on the test results, we select and present the best sentiment classification algorithm for airline services companies.

Our work can benefit researchers and decision makers in airline services companies studying customers' feedback and satisfaction for their companies' services. Researchers and decision makers in airline services companies can utilize the Bayesian Network algorithm to automatically classify customers' feedback on micro-blogging platforms like Twitter. Business analysis applications can be developed from the Bayesian Network algorithm as well.

In the next section, we discuss some related works on sentiment analysis and give an outline of current studies in twitter sentiment analysis area. Then the data collection

process is described and data features are explained. The following section displays algorithms including the lexicon-based method, the SVM method, the Naive Bayesian method and the Bayesian Network method. The experiment section describes the experimental results from different sentiment classification algorithms tested on airline services datasets. Finally, we conclude this paper by presenting the best sentiment analysis algorithm for airline services and suggest several directions of future work.

2 Related Work

Sentiment mining is a division of text mining, which includes information retrieval, lexical analysis and many other techniques. Many methods widely applied in text mining are exploited in sentiment mining as well. But the special characters of sentiment expression in language make it very different from standard factual-based textual analysis. The most important application of opinion mining and sentiment analysis has been customer review mining. There have been many studies recorded on different review sites.

Li, Feng and Xiao used a multi-knowledge based approach in mining movie reviews and summarizing sentiments, which proved very effective in applications [1]. Ding, Bing and Philip proposed a holistic lexicon based approach to classify customer' sentiments towards certain products and achieved high accuracy [2]. This approach is content dependent and needs to select feature words, phrases from training data. Lin and He proposed a probabilistic modeling framework called Joint-sentiment model, which adopted the unsupervised machine learning method [3]. In their research, they applied their model in movie reviews and classify the review sentimental polarity. Prabowo and Thelwall combine ruled-based classification, supervised learning and machine learning methods and proposed a hybrid method. Their method yielded satisfactory results when applied to movie reviews, product reviews and Myspace comments [4]. In the research of Wilson et al, they exploited hashtags in tweets to spot tweets, which can be used as training data. 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]. Beside that, Wilson et al also considered three polarities in tweets sentiment classification, which includes positive sentiment, negative sentiment and neutral sentiment. Unigrams, bigrams and POS features were taken into account as classification features, and emoticons and other non-textual features were also considered. In their experiments, it showed that training data with hashtags could train better classifiers than regular training data. But in their research, the dataset were from libraries and they neglected the fact hashtagged tweets are only a small part of real world tweets dataset.

Pak and Paroubek proposed an approach, which can retrieve sentimental oriented tweets from twitter API and classify their sentiment orientations [6]. From the test result, they found that bi-gram term classifier produced highest classification accuracy because it achieves a good balance between coverage and precision. Their work in tweets sentiment mining is not domain specific, which means applying their methods

in domain specific mining will yield different results. And the data source is biased as well because they retrieved only the tweets with emoticons and neglected all other tweets that didn't contain emoticons, which is the majority of tweet data. In this work, they didn't consider the existence of neutral sentimental tweets and classifying those tweets from the sentimental tweets is very important for tweet analysis.

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, Jet Blue Airlines and Southwest Airlines. They adopted conventional text analysis methods to study twitter user's interactions and provided advice for airline companies for micro-blogging campaign. In their research, they didn't 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, this method was applied and tested with pre-labeled data. It yielded inaccurate testing results because sentiment classifications are highly domain specific.

Adeborna et al adopted Correlated Topics Models (CTM) with Variational Expectation-Maximization (VEM) algorithm [9]. Their lexicons for classification were developed with AQR criteria. In Sentiment detection process, SVM, Entropy and Naive Bayesian were compared and Naive bays method was adopted. Besides that, tweets are categorized by topics using CTM with the VEM algorithm. The result of this case study reached 86.4% accuracy in subjectivity classification and displayed specific topics describing the nature of the sentiment. In this research, the author only used unigrams as sentiment classification features in Naive Bayes algorithm, which can cause problems because phrases and negation terms can change sentiment orientation of those terms in sentences. In my work, Unigrams, Bigrams and the information gain algorithm will be applied into feature selections, which yields higher accuracy. Besides that, their work did not present details about the classification approaches and comprehensive evaluation. However, our work focus on the comparison of the performances of different approaches and we gave a detail evaluation of those approaches.

3 Data Preparation

We connected to Twitter Search API and retrieved tweets, which contained key words including airline brands and the word "flight". Tweets retrieved from Twitter Search API in this way fully meet the expectation and generate little ambiguity. In this section we outline our data preparation process and describe the attributes of our dataset.

To get a full and comprehensive coverage of tweets about North American airline services, most of the airline services brands in North America were considered. Based on the list, the largest airlines in North America are: Delta Airline, Jetblue Airline, United Airline, AirCanada, Southwest Airline, Airtran Airline, Westjet Airline, American Airline, Frontier Airline, Virgin Airline, Allegiant Airline, Spirit Airline, US Airline, Hawaiian Airline, Skywest Airline, Alaska Airline[10]. Retrieving tweets about those brands can build the best dataset for sentiment analysis of airline services.

Using Twitter Search API to retrieve tweets by key words might cause ambiguity. For example, searching tweets with the key word 'Delta', which is the biggest airline brand in North America, might collect tweets that convey geographic information other than Delta airline services feedback. In our work, we search each airline brand with a combination of two key words including the brand's name and 'flight' to collect tweets that convey airline services feedback. In the process the labeling tweets, the irrelevant tweets, which were retrieved caused by the ambiguity of query, were discarded.

For sentiment analysis, only the text of tweets was considered; there was no other constraint for retrieved tweets except the language is set to English. We retrieved tweets with those sixteen brands' names and the key word 'flight' from Twitter Search API. However, Twitter Search API only returns 3000 tweets in maximum and 200 tweets in minimum for a single query each time. Because timing factors were not considered in our work, we kept retrieving tweets randomly in different periods until the data volume meets our requirement. At the end, we got 8086 tweets for Delta Airlines, 5060 tweets for United Airlines, 4800 for Southwest Airlines, 6000 tweets for AirCanada and 3807 tweets for Jetblue Airlines and 4135 for rest of airline companies. Because the volume of tweets returned from Twitter Search API for each brand indicates that its market share, the fractions of tweets for each brands were not adjusted. In total, there was a dataset containing 31888 tweets in our work.

These tweets include original tweets and retweets. We discard the irrelevant tweets and labeled each relevant tweet in the dataset as positive sentiment, negative sentiment or neutral sentiment manually. In the dataset, 2502 tweets were labeled positive, 7039 tweets were labeled negative, 13074 tweets were labeled neutral and 9273 tweets were discard for being irrelevant.

Table 1. Tweet class distribution

class	positive	negative	neutral	irrelevant
tweets	2502	7039	13074	9273

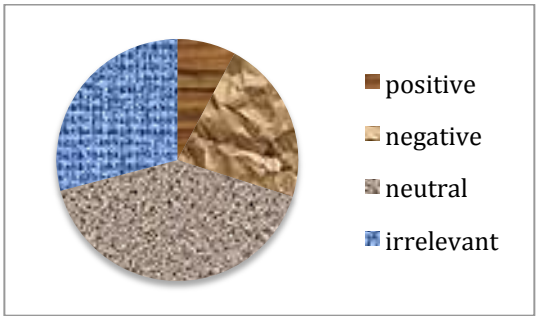

Figure 1. Tweet class distribution

Labeled tweets were used to train classifiers by supervised learning methods and used to test classification as well. In the Bayesian approaches, model training process requires the class distribution to be balanced. So we resample the data with 2500 tweets for each class: positive sentiment, negative sentiment and neutral sentiment. For evaluation purpose, the dataset with 7500 tweets was used for every classification approach in our experiment.
In our work, we removed all symbols, hashtag signs, links, emoticons and punctuations from tweets since we don't regard those factors as classification features. We also adopted text clean techniques by using the tm package in R to remove duplicates and clean tweet. We used Weka as our data mining tool to implement our experiment.

4 Classification Approaches

Here we describe four different classifiers using different classification methods. They are the Lexicon-based classifier, SVM classifier, Naive Bayes classifier and Bayesian Network classifier.

4.1 Lexicon-based classifier

This classifier is not constructed by machine learning. In this method, two sentiment word lists are utilized to score each tweet document and determine its sentiment orientation. This method treats each tweet document as a bag-of-words and doesn't take semantic structures into consideration. The lexicon-based classifier passes each tweet document and matches them with the positive word list and the negative word list. The occurrences of matches are scored and the final score for each tweet document is the result of positive scores minus negative scores. If the result is bigger than 0, the tweet is classified as positive, and if the result is less than 0, the tweet is classified as negative. Otherwise, if the result is equal to 0, the tweet is classified as neutral.
In our work, we adopted the word lists produced by Hu and Liu in their work Mining and Summarizing Customer Reviews [11] and we added four words including 'de-

layed', 'late', 'oversold' and 'bumped' into the negative word list because those words indicate strong negative sentiment in the airline services domain.

4.2 Naive Bayesian Classifier

The Naïve Bayesian method is one of the most widely used methods to classify text data. The Naïve 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. Like the lexicon-based classifier, the Naïve Bayesian classifier treats each tweet document as a bag-of-words. In our work, we calculate the sentimental orientation probabilities based on the Naïve Bayesian algorithm for each word occurring in the training dataset and set up the sentiment distribution metrics for all of words in the training dataset.

$$p(S|D) = \frac{p(S)}{p(D)} \prod_{i}^{n} p(w_i|S) \quad (1)$$

The probability of each tweet document for each of the three sentiment categories is calculated as shown in formula 1. p(S | D) represents the probability of document D being classified as sentiment category S. p(S) represents the probability of sentiment category S and p(D) represents the probability of document D. p(w_i | S) represents the probability of occurrence of word w_i in sentiment S. The number n represents the total number of words for document D. The Naïve Bayesian classifier passes a single tweet document and calculates the products of the probabilities of every word 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 Weka to implement experiments and tests.

4.3 SVM Classifier

Support vector machine classifiers are supervised machine learning models used for binary classification and regression analysis. However, in our work, we aim to build classifiers, which can classify tweets into three sentiment categories. Based on the study done by Hsu and Lin, the pairwise classification method outperforms the one-against-all classification method in multiclass support vector machine classification [12]. In the pairwise classification method, each pair of classes will have one SVM

trained to separate the classes. The accuracy of the classification will be the overall accuracy of every SVM classification included.
We adopted pairwise classification approach in the SVM classification method. We utilized the SMO algorithm in Weka, which use pairwise classification for multiclass SVM classification, in Weka to train the SVM classifier and implement experiments and tests.

4.4 Bayesian Network Classifier

Like Naïve Bayesian method, Bayesian Network also derives from Bayes' theorem [13], but Naïve Bayesian method assumes that the features are independent to each other. However, Bayesian Network method takes consideration of the relationships between the features.

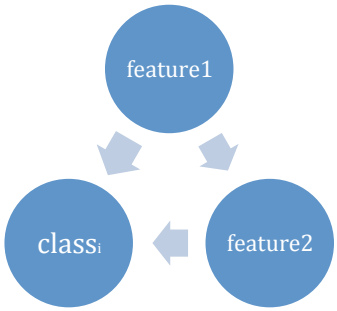

Figure 2. Bayesian Network model

As illustrated above, feature1 and feature2 are the features which decides the probability of $class_i$, but the occurrence of the feature1 influences the occurrence of feature 2, which means the two features are not independent. The Bayesian Network algorithm can be described with the formula below:

$$p(class_i) = \prod_{f \in F} p\left(f \middle| pa(f)\right) \quad (2)$$

In formula 2, $p(class_i)$ represents the probability for the instance being classified as $class_i$. $p(f|pa(f))$ represents the probability of feature f given their parent features $pa(f)$. F represents the feature set. The Bayesian Network classifier passes each single tweet can calculates the probability for each class: positive, negative and neutral. Each tweet will be classified as the class which gets the highest probability.

5. Experiments

We conducted experiments with the four classification models. We used the 10 folds validation plan to evaluate the machine learning classification approaches includes: the Naïve Bayesian classifier, the SVM classifier and the Bayesian Network classifier. All of our conductions were implemented in R and Weka. Test results for three class classification experiment are shown in table 4. The lexicon-based classifier gets the lowest accuracy, which is 60.5%. The accuracy of the Bayesian Network model classification reached 85.1%. The Naïve Bayesian classifier outperformed the lexicon-based classifier and the Naïve Bayesian classifier by reaching an accuracy of 85.3%. The Bayesian Network classifier produced the highest accuracy and reached 87.2%

Table 2. Accuracy of 3 class classification

Classifier	Positive accuracy	Negative Accuracy	Neutral Accuracy	Overall Accuracy
Lexicon-based Classifier	70.8%	56.2%	54.6%	60.5%
Naïve Bayesian Classifier	86.2%	84.0%	84.0%	85.3%
Bayesian Network Classifier	86.5%	83.4%	85.3%	85.1%
SVM Classifier	86.9%	89.1%	85.5%	87.2%

Besides, we also implemented the sentiment classification algorithms in the two polarity classification experiment, in which the training data and the test data only contain two classes: positive sentiment and negative sentiment. In our experiment, the accuracy of the Lexicon based classifier is 67.9%, the accuracy of Naïve Bayesian classifier is 91.3%, the accuracy of Bayesian Network classifier is 91.4% and the accuracy of SVM classifier is 91.9%. The results shows that, the sentiment classification algorithms perform better in two class classification than in three class classification.

Table3. Accuracy of binary class classification

Classifier	Positive accuracy	Negative Accuracy	Overall Accuracy
Lexicon-based Classifier	77.8%	58.0%	67.9%
Naïve Bayesian Classifier	91.2%	91.3%	91.3%
Bayesian Network Classifier	91.4%	91.5%	91.4%

SVM Classifier	91.9%	91.8%	91.9%

For the Lexicon based classifier, in both of the two class classification and three class classification experiments, the positive accuracies are much higher than the negative accuracies. That is because many Twitter users tweet their feelings in ironic ways, in which positive words are used to express negative feelings. In both experiments, the SVM classifier produced the highest accuracies, which indicates that SVM algorithm will be the most suitable sentiment classification methods for tweets about airline services.

6. Conclusions

We have compared four classification methods for Twitter sentiments of airline services. We build four classifiers and selected the best sentiment classifier, which is the SVM classifier. This classifier can be used for airline services business analysis applications, which will be able to automatically classify customer's satisfaction about airline services. We identify several directions for future work. First, we discovered that the negative and positive feedbacks from customers' tweets were about several different topics, like oversold problems and delay issues. We plan to work on this domain to combine topic recognition methods and the SVM sentiment classification for airline services. Besides, we will extend the research to generalizing the sentiment classification model training approach for those domain dependent tweets. In the end, we also like to discover some rules and knowledge in twitter sentiment classification, which can further improve the sentiment classification accuracy.

References

1. Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. "Movie review mining and summarization." Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006.
2. Ding, Xiaowen, Bing Liu, and Philip S. Yu. "A holistic lexicon-based approach to opinion mining." *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008.
3. Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009
4. Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." *Journal of Informetrics* 3.2 (2009): 143-157.
5. Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." *ICWSM* 11 (2011): 538-541.
6. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. 2010.

7. Sreenivasan, Nirupama Dharmavaram, Chei Sian Lee, and Dion Hoe-Lian Goh. "Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines." *Program: electronic library and information systems* 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. Adeborna, Esi, and Keng Siau. "AN APPROACH TO SENTIMENT ANALYSIS–THE CASE OF AIRLINE QUALITY RATING." (2014).
10. Wikipedia contributors. "List of largest airlines in North America." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 12 Oct. 2014. Web. 15 Oct. 2014.
11. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
12. Hsu, Chih-Wei, and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." *Neural Networks, IEEE Transactions on* 13.2 (2002): 415-425.
13. Wikipedia contributors. "Bayes' theorem." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia*, 18 Jan*. 2015. Web. *21 Jan*. 2015.
14. Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.